UNIVERSIDAD NACIONAL DEL SANTA

ESCUELA DE POSGRADO

Programa de Maestría en Ingeniería de Sistemas e Informática mención Gestión de Tecnologías de Información

Minería de datos usando el algoritmo de Clustering K-Means aplicado a las facturas electrónicas de la empresa Envases Los Pinos S.A.C., 2023

Tesis para optar el grado de Maestro en Ingeniería de Sistemas e Informática mención Gestión de Tecnologías de la Información

Autor:

Bach. Ponte Arica, Anthony Rosemberg

Código ORCID: 0000-0002-7146-6700

Asesor:

Ms. Manrique Ronceros, Mirko Martin

DNI Nº 32965599

Código ORCID: 0000-0002-0364-4237

Nuevo Chimbote - PERÚ

2026

UNS
ESCUELA DE POSGRADO

CERTIFICACIÓN DEL ASESOR

Yo, Ms. Mirko Martin Manrique Ronceros, certifico mi asesoramiento de la tesis titulada: **Minería de datos usando el algoritmo de Clustering K-Means aplicado a las facturas electrónicas de la empresa Envases Los Pinos S.A.C., 2023,** que tiene como autor a **Ponte Arica, Anthony Rosemberg,** alumno de la **Maestría en Ingeniería de Sistemas e Informática, mención Gestión de Tecnologías de la Información**, ha sido elaborado de acuerdo al Reglamento General de Grados y Títulos de la Universidad Nacional del Santa.

Ms. Mirko Martin Manrique Ronceros
Asesor
Código ORCID: 0000-0002-0364-4237
DNI N° 32965599

AVAL DEL JURADO EVALUADOR

Tesis titulada: **Minería de datos usando el algoritmo de Clustering K-Means aplicado a las facturas electrónicas de la empresa Envases Los Pinos S.A.C., 2023,** que tiene como autor a **Ponte Arica, Anthony Rosemberg.**

Tesis para optar el grado de Maestro en Ingeniería de Sistemas e Informática mención Gestión de Tecnologías de la Información

Revisado y Aprobado por el Jurado Evaluador:

Ms. Manco Pulido, Pedro Glicerio
Presidente
DNI. N° 32953190
Código ORCID: 0000-0002-8542-2119

Ms. Gil Narvaez, Carlos Alfredo
Secretario
DNI. N° 32970648
Código ORCID: 0000-0003-0137-0444

Ms. Mirko Martin Manrique Ronceros
Vocal/Asesor
DNI N° 32965599
Código ORCID: 0000-0002-0364-4237

UNS
ESCUELA DE POSGRADO

ACTA DE EVALUACIÓN DE SUSTENTACIÓN DE TESIS

A los siete días del mes de enero del año 2026, siendo las 12:00 horas, en el aula P-01 de la Escuela de Posgrado de la Universidad Nacional del Santa, se reunieron los miembros del Jurado Evaluador, designados mediante Resolución Directoral Nº 962-2025-EPG-UNS de fecha 28.11.2025, conformado por los docentes: Ms. Pedro Glicerio Manco Pulido (Presidente), Ms. Carlos Alfredo Gil Narvaez (Secretario) y Ms. Mirko Martin Manrique Ronceros (Vocal); con la finalidad de evaluar la tesis intitulada: "***MINERÍA DE DATOS USANDO EL ALGORITMO DE CLUSTERING K-MEANS APLICADO A LAS FACTURAS ELECTRÓNICAS DE LA EMPRESA ENVASES LOS PINOS S.A.C., 2023***"; presentado por el tesista **Anthony Rosemberg Ponte Arica,** egresado del programa de Maestría en Ingeniería de Sistemas e Informática Mención Gestión de Tecnología de Información.

Sustentación autorizada mediante Resolución Directoral Nº 006-2026-EPG-UNS de fecha 05 de enero de 2026.

El presidente del jurado autorizó el inicio del acto académico; producido y concluido el acto de sustentación de tesis, los miembros del jurado procedieron a la evaluación respectiva, haciendo una serie de preguntas y recomendaciones al tesista, quien dio respuestas a las interrogantes y observaciones.

El jurado después de deliberar sobre aspectos relacionados con el trabajo, contenido y sustentación del mismo y con las sugerencias pertinentes, declara la sustentación como BUENO, asignándole la calificación de DIECIOCHO.

Siendo las 12:45 horas del mismo día se da por finalizado el acto académico, firmando la presente acta en señal de conformidad.

Ms. Pedro Glicerio Manco Pulido
Presidente

Ms. Carlos Alfredo Gil Narvaez
Secretario

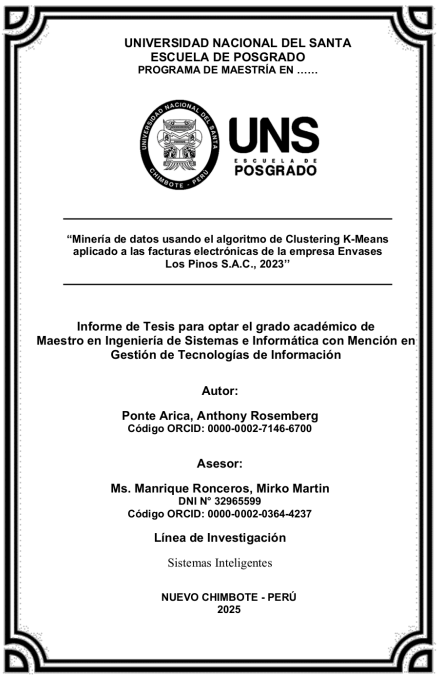
Ms. Mirko Martin Manrique Ronceros
Vocal/Asesor

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	Anthony Ponte
Título del ejercicio:	Tesis
Título de la entrega:	TESIS-PONTE_ANTHONY.pdf
Nombre del archivo:	TESIS-PONTE_ANTHONY.pdf
Tamaño del archivo:	3.15M
Total páginas:	123
Total de palabras:	24,070
Total de caracteres:	144,026
Fecha de entrega:	06-feb-2026 03:28p. m. (UTC-0500)
Identificador de la entrega:	2872776234



UNIVERSIDAD NACIONAL DEL SANTA
ESCUELA DE POSGRADO
PROGRAMA DE MAESTRÍA EN .....

UNS
ESCUELA DE POSGRADO

"Minería de datos usando el algoritmo de Clustering K-Means aplicado a las facturas electrónicas de la empresa Envases Los Pinos S.A.C., 2023"

Informe de Tesis para optar el grado académico de Maestro en Ingeniería de Sistemas e Informática con Mención en Gestión de Tecnologías de Información

Autor:

Ponte Arica, Anthony Rosemberg
Código ORCID: 0000-0002-7146-6700

Asesor:

Ms. Manrique Ronceros, Mirko Martin
DNI N° 32965599
Código ORCID: 0000-0002-0364-4237

Línea de Investigación

Sistemas Inteligentes

NUEVO CHIMBOTE - PERÚ
2025

TESIS-PONTE_ANTHONY.pdf

INFORME DE ORIGINALIDAD

15%	13%	2%	7%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.unac.edu.pe Fuente de Internet	1%
2	idoc.pub Fuente de Internet	1%
3	repositorio.uns.edu.pe Fuente de Internet	1%
4	repositorio.unc.edu.pe Fuente de Internet	<1%
5	hdl.handle.net Fuente de Internet	<1%
6	Submitted to Universidad Nacional del Santa Trabajo del estudiante	<1%
7	www.coursehero.com Fuente de Internet	<1%
8	Submitted to Universidad Cesar Vallejo Trabajo del estudiante	<1%
9	Submitted to Universidad Internacional Isabel I de Castilla	<1%

DEDICATORIA

A Dios, por darme fortaleza en los momentos de duda. Sin Su luz, este camino no habría sido posible.

A mis padres Jesus Maria Arica Valencia y Maximo Rosemberg Ponte Angeles, por enseñarme con el ejemplo el valor del esfuerzo, la honestidad y la perseverancia. Gracias por creer en mí.

A mi novia Adele Cristi Paredes Villanueva, por estar a mi lado en cada etapa de este proceso, por su apoyo emocional, su paciencia y sus palabras de aliento que me impulsaron a seguir adelante.

A mi familia, por cada gesto de apoyo y por ser ese pilar silencioso que siempre sostiene sin pedir nada a cambio.

AGRADECIMIENTOS

En primer lugar, agradezco a Dios por darme la fortaleza, la salud y la perseverancia para completar esta etapa de mi vida.

A mis padres, por su apoyo incondicional y sus palabras de aliento. Su ejemplo de esfuerzo y dedicación me impulsó a seguir adelante.

A mi asesor de tesis, Mirko Manrique, por su orientación, paciencia y valiosos aportes durante todo el proceso. Su experiencia y compromiso fueron fundamentales para el desarrollo de este trabajo.

A los docentes del programa de Maestría, por compartir sus conocimientos y por contribuir a mi formación profesional y humana.

A compañeros de estudio, con quienes compartí largas horas de trabajo, discusiones académicas y también momentos de alegría y camaradería. Este logro también es suyo.

Finalmente, agradezco a Universidad Nacional del Santa, por brindarme el espacio, los recursos y el respaldo necesario para llevar a cabo esta investigación

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

La presente investigación se desarrolló en la empresa Envases Los Pinos S.A.C. durante el año 2023, en respuesta a la creciente necesidad de optimizar los procesos de gestión administrativa y toma de decisiones contables, debido al elevado volumen de facturación electrónica y la falta de herramientas analíticas que permitan una adecuada segmentación y análisis de la información. En este contexto, se identificó como problema central la limitada capacidad para organizar, interpretar y aprovechar estratégicamente los datos contenidos en las facturas electrónicas, lo que repercutía en la eficiencia operativa y en la oportunidad de las decisiones gerenciales.

El objetivo general del estudio fue determinar la influencia de la minería de datos mediante el algoritmo de clustering K-Means en la mejora de la gestión administrativa y en la toma de decisiones estratégicas basadas en la información de las facturas electrónicas. Para ello, se adoptó un enfoque cuantitativo, de tipo aplicado y con diseño preexperimental, utilizando una muestra de 1250 facturas agrupadas en 25 lotes de 50 unidades cada uno. Se realizaron mediciones pre y post implementación del modelo para comparar los resultados sobre distintos indicadores claves.

Los resultados evidenciaron mejoras sustanciales en todos los indicadores. El tiempo promedio en la clasificación de facturas se redujo en un 82.34 %, mientras que la exactitud en los reportes se incrementó de 54.87 % a 99.15 %. Asimismo, la eficiencia en la detección de patrones pasó de 18.96 % a 91.22 %, y la tasa de mejora en la toma de decisiones mostró una reducción promedio de 238.03 minutos, equivalente al 80.07 %. Todos los resultados fueron estadísticamente significativos ($p < 0.001$), con tamaños del efecto altos (Cohen's $d > 2.0$) y niveles de fiabilidad aceptables ($\alpha > 0.75$).

Se concluyó que la aplicación del algoritmo K-Means influyó de manera positiva y significativa en la mejora de los procesos administrativos y contables de la empresa, al proporcionar herramientas de segmentación automatizada que optimizaron el tiempo, incrementaron la exactitud y facilitaron la toma de decisiones basadas en datos. Estos hallazgos respaldan la viabilidad de incorporar técnicas de minería de datos en entornos empresariales con alto flujo documental y necesidades analíticas avanzadas.

Palabras Claves: Minería de datos, Clustering K-Means, Facturación Electrónica, Toma de decisiones, Gestión administrativa.

ABSTRACT

This research was conducted at Envases Los Pinos S.A.C. during 2023, in response to the growing need to optimize administrative management processes and accounting decision-making due to the high volume of electronic invoicing and the lack of analytical tools that allow for proper segmentation and analysis of information. In this context, the central problem identified was the limited capacity to organize, interpret, and strategically leverage data contained in electronic invoices, which affected operational efficiency and the timeliness of managerial decisions.

The main objective of the study was to determine the influence of data mining using the K-Means clustering algorithm on improving administrative management and strategic decision-making based on electronic invoice data. An applied, quantitative approach with a pre-experimental design was adopted, using a sample of 1,250 invoices grouped into 25 batches of 50 units each. Pre- and post-implementation measurements of the model were carried out to compare results across several key indicators.

The results showed significant improvements across all indicators. The average time for invoice classification was reduced by 82.34%, while report accuracy increased from 54.87% to 99.15%. Additionally, pattern detection efficiency jumped from 18.96% to 91.22%, and the decision-making improvement rate showed an average reduction of 238.03 minutes, equivalent to 80.07%. All findings were statistically significant ($p < 0.001$), with large effect sizes (Cohen's $d > 2.0$) and acceptable reliability levels ($\alpha > 0.75$).

It was concluded that the application of the K-Means algorithm positively and significantly influenced the improvement of the company's administrative and accounting processes by providing tools for automated segmentation that optimized time, increased accuracy, and facilitated data-driven decision-making. These findings support the viability of incorporating data mining techniques in business environments with high document flow and advanced analytical needs.

Keywords: Data mining, K-Means Clustering, Electronic invoicing, Decision making, Administrative management.

I. INTRODUCCIÓN

1.1. Descripción del Problema

1.1.1. Realidad Problemática

En un mundo globalizado, el comercio internacional ha experimentado una transformación digital acelerada, impulsada por la adopción de tecnologías de la información y la comunicación.

En el contexto internacional; en Asia, la adopción de la facturación electrónica se había consolidado como una de las estrategias clave para la transformación digital y la transparencia fiscal en los últimos años. Japón, por ejemplo, implementó en 2023 el sistema de facturación electrónica calificada con el objetivo de garantizar un control tributario más preciso y reducir la evasión fiscal, generando una carga administrativa significativa para las pequeñas y medianas empresas, pero también oportunidades de innovación en los procesos financieros (Yanagawa, 2023). Malasia, por su parte, planificó la implementación obligatoria de la facturación electrónica a partir de junio de 2024, desarrollando un marco adaptativo para garantizar la transición digital en sectores empresariales heterogéneos, con el fin de fortalecer la eficiencia económica y la transparencia fiscal (Hong & Shibghatullah, 2024). Estas iniciativas reflejaban una tendencia creciente en Asia hacia la digitalización de los procesos administrativos y financieros, que impactaban de manera directa en la competitividad empresarial.

Las estadísticas mostraban que en la región Asia-Pacífico más del 60 % de las grandes corporaciones habían adoptado sistemas de facturación electrónica antes de 2023, mientras que las pequeñas y medianas empresas avanzaban de manera más lenta debido a limitaciones técnicas y de inversión (Yanagawa, 2023). En países como Corea del Sur, la obligatoriedad de la facturación electrónica había logrado reducir hasta en un 30 % los costos operativos de las empresas que la implementaron plenamente (Hong & Shibghatullah, 2024). Sin embargo, este proceso no estuvo exento de desafíos: la falta de interoperabilidad entre plataformas, los costos de adaptación tecnológica y las brechas de capacitación se presentaban como obstáculos persistentes. Estos aspectos evidenciaban que, si bien Asia se encontraba a la vanguardia en la adopción de la facturación

electrónica, aún existía un amplio margen de mejora en términos de integración tecnológica y aprovechamiento de herramientas avanzadas como la minería de datos aplicada a grandes volúmenes de información financiera.

En Europa, la facturación electrónica había seguido un camino marcado por las regulaciones comunitarias y la estandarización de procesos. Desde la Directiva 2014/55/UE, los Estados miembros habían avanzado en la obligatoriedad de la facturación electrónica en el sector público, y países como Italia y Polonia establecieron su uso obligatorio en el ámbito B2B para 2019 y 2024, respectivamente (Selera, 2023). Este marco normativo había impulsado la digitalización en sectores económicos estratégicos, promoviendo ahorros sustanciales y mayor eficiencia en la gestión tributaria. Según estimaciones de la Comisión Europea, la adopción generalizada de la facturación electrónica en la Unión Europea generaba un ahorro potencial de hasta 40.000 millones de euros anuales (Bojanc, Pucihar & Lenart, 2024).

No obstante, el nivel de implementación no había sido homogéneo en toda Europa. Mientras que países del norte y occidente mostraban una alta adopción, en naciones con menor capacidad tecnológica la penetración era reducida, lo cual limitaba los beneficios de la interoperabilidad transfronteriza (Šoltésová, 2022). Además, las pequeñas y medianas empresas continuaban enfrentando barreras vinculadas a los costos iniciales de adopción y a la complejidad de los estándares técnicos exigidos. En este contexto, diversos autores resaltaron que la facturación electrónica en Europa no solo representaba una obligación normativa, sino también una oportunidad para la automatización de procesos y la utilización de técnicas avanzadas de análisis de datos, las cuales podían convertirse en un catalizador de innovación y sostenibilidad empresarial (Bojanc et al., 2024).

En Latinoamérica se había observado un crecimiento significativo en la adopción de la facturación electrónica impulsado por reformas tributarias orientadas a mejorar la transparencia fiscal y reducir la evasión. En Ecuador, por ejemplo, el sistema de facturación electrónica fue lanzado oficialmente en 2013 como estrategia para incrementar el cumplimiento del Impuesto al

Valor Agregado (IVA), y entre 2014 y 2016 había evidencias de que la declaración de ventas, compras y obligaciones tributarias aumentó conforme se expandió la cobertura del sistema electrónico (Ramírez-Álvarez, Oliva & Andino, 2022). En ese mismo país, un estudio reciente que abarcó el periodo 2019-2023 reportó que el uso obligatorio de facturación electrónica contribuyó a una reducción tangible de la evasión fiscal en el sector comercial de pequeñas y medianas empresas, aunque dicha reducción estuvo condicionada por barreras como la informalidad, falta de capacitación y déficits en los controles tributarios (Uquillas Granizo & López Naranjo, 2024). En Chile y Colombia, se registraron avances normativos y técnicos que permitieron que las administraciones tributarias mejoraran su capacidad para monitorear transacciones tributarias mediante la facturación electrónica. En estos países se consiguieron mejoras en la recaudación de impuestos correlacionadas con una mayor digitalización de los procesos tributarios y mayor cobertura del sistema (Rodríguez, 2023; Hernandez Aros et al., 2023). No obstante, a pesar de estos logros, persistían desafíos significativos en la región: muchas pequeñas y medianas empresas enfrentaban dificultades para adaptarse a los requerimientos técnicos del sistema, los costos de implementación resultaban elevados para empresas de menor escala, existían lagunas en interoperabilidad de plataformas tributarias, y en varios países la informalidad seguía limitando los efectos esperados de la facturación electrónica (Uquillas Granizo & López Naranjo, 2024; Hernandez Aros et al., 2023). Además, aunque algunos estudios cuantificaron efectos, pocos exploraron el comportamiento interno de los datos de las facturas electrónicas: es decir, cómo patrones de montos, frecuencia, productos o clientes podrían agruparse o presentar anomalías que aportaran insights adicionales para la toma de decisiones empresariales y fiscales.

En el contexto nacional, la facturación electrónica se había convertido en un eje estratégico de modernización tributaria impulsado por la Superintendencia Nacional de Administración Tributaria (SUNAT). A través de la Resolución de Superintendencia N.º 128-2021/SUNAT, se estableció que a partir del 1 de junio de 2022 la facturación electrónica sería

obligatoria para todos los contribuyentes, cerrando así el último grupo que todavía operaba con comprobantes físicos (Sovos, 2022). Este mandato generó un cambio estructural en el cumplimiento fiscal de miles de empresas, desde grandes corporaciones hasta micro y pequeñas empresas (MYPE), obligándolas a digitalizar sus procesos contables y tributarios.

Sin embargo, el proceso no estuvo exento de dificultades. En Lima Metropolitana, un estudio evidenció que entre 2019 y 2022 muchas MYPE todavía presentaban incumplimientos parciales en sus obligaciones tributarias debido a la tardía implementación del Sistema de Emisión Electrónica (SEE) y de los libros electrónicos, lo que generaba errores administrativos y mayores costos operativos (Puican Núñez & Sánchez Herrada, 2024). En regiones como Puno, específicamente en la ciudad de Juliaca, se identificó que el 83 % de los empresarios del sector textil reconocía que la facturación electrónica representaba un mecanismo formal de control tributario, y un 96 % afirmaba que facilitaba el cumplimiento contable; sin embargo, las limitaciones tecnológicas y la resistencia al cambio constituían barreras significativas para su adopción plena (Becerra Paredes, 2024).

La relación entre facturación electrónica y obligaciones tributarias también había sido analizada en sectores específicos. En una empresa médica, se halló una correlación positiva significativa ($r = 0,695$) entre el uso de comprobantes electrónicos y el cumplimiento fiscal durante el año 2022, demostrando que la digitalización favorecía la formalización de los contribuyentes (Cubas Burgos, 2022). A pesar de estos resultados, en la región Piura se reportó que muchas empresas no habían percibido un impacto económico favorable inmediato de la facturación electrónica, debido a los costos de implementación y la carencia de soporte técnico especializado, lo que limitaba la eficiencia esperada de la medida (Universidad Nacional de Piura, 2022). En conjunto, los estudios nacionales mostraban que la facturación electrónica había fortalecido el control tributario y la transparencia en el Perú. No obstante, la explotación analítica de los datos generados por estas facturas aún era limitada. La información contenida en millones de comprobantes electrónicos permanecía

subutilizada, ya que se enfocaba únicamente en el cumplimiento normativo y no en el análisis estratégico de los datos.

En el contexto local, en la región Áncash la modernización de los procesos administrativos y tributarios de las empresas había representado un desafío constante, especialmente en el marco de la obligatoriedad de la facturación electrónica impuesta por la SUNAT a nivel nacional desde 2022. A pesar de los avances en la implementación de sistemas digitales, se había observado que muchas empresas enfrentaban demoras sustanciales en sus procesos administrativos relacionados con la emisión, gestión y control de comprobantes de pago, especialmente al migrar hacia facturación electrónica o sistemas informáticos de facturación. A pesar de estos avances puntuales, no se había investigado con profundidad cómo se comportaban los datos generados por las facturas electrónicas en empresas de Áncash: qué patrones de frecuencia, montos, productos o clientes emergían de dichos datos, ni se había aplicado algoritmos de minería de datos como clustering (por ejemplo, K-Means) para descubrir agrupaciones útiles para optimizar la gestión administrativa, fiscal, contable o comercial. En muchos casos los estudios se limitaron a medir mejoras de tiempos, costos o satisfacción, sin adentrarse en un análisis más granular de los datos transaccionales. Asimismo, no se encontraron estadísticas específicas para empresas de la región Áncash que detallaran el volumen de facturas electrónicas emitidas por tipo de contribuyente, ni cuantificaciones de ahorros operativos expresados en porcentaje generalizado para la región. En contraste con otras regiones del país donde estudios como los de Juliaca en la ciudad homónima mostraron influencia significativa de la facturación electrónica en el tratamiento contable (Becerra Paredes, 2024), en Áncash persistía una carencia de investigaciones que aplicaran técnicas de minería de datos sobre las facturas electrónicas para identificar anomalías, segmentos de clientes, productos o patrones de compra/venta, lo cual limitaba el aprovechamiento estratégico de la información generada por los sistemas implementados.

1.1.2. Análisis del Problema

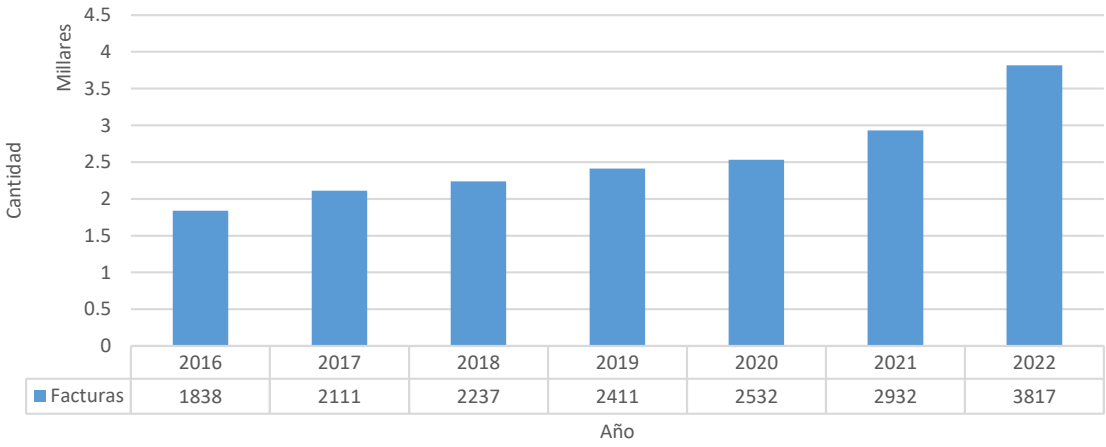
EPINSA (Envases Los Pinos S.A.C., 2023) es una empresa que forma parte del Grupo COMECA de Costa Rica el cual cuenta con 70 empresas alrededor de América Latina participando en 4 sectores principalmente: papel, cartón, envases (de hojalata, plástico y flexible) y supermercados. EPINSA atiende al mercado peruano desde el año 2006, consolidado actualmente en la venta de envases de hojalata para los sectores Pesca y Agro e incrementando su participación en exportaciones hacia Ecuador, Colombia y Chile.

Como persona jurídica EPINSA fue incorporado al Régimen de Buenos Contribuyentes a partir del 01/08/2015 por la Resolución N° 0110050001045 emitida por la Superintendencia Nacional de Aduanas y de Administración Tributaria (SUNAT). Y desde el año 2016 a estado obligada a emitir comprobantes de pago electrónicos.

En la Figura 1: Facturas emitidas Envases Los Pinos S.A.C. – 2016 - 2020 podemos ver la progresión de la emisión de comprobantes electrónicos dentro de la empresa del 2016 al 2022:

Figura 1:

Facturas emitidas Envases Los Pinos S.A.C. – 2016 - 2020



Nota. La figura muestra las cifras de facturas emitidas del año 2016 hasta el 2022.

Podemos ver que el aumento de emisión de facturas ha ido incrementando del 2016 al 2022, llegando a ser abrumadora y difícil de entender lo que ha complicado el hallazgo de patrones significativos y características comunes de los clientes, reduciendo la precisión del proceso de segmentación de clientes. Esto degrada la eficacia de la toma de decisiones empresariales estratégicas, tales como:

- Diseño de una experiencia personalizada para cada grupo clientes.
- Diseño de estrategias para retención de clientes.
- Diseño de incentivos para aumentar las ventas manteniendo la lealtad de los clientes.

La empresa Envases Los Pinos S.A.C., ubicada en la región Áncash, se dedicaba a la producción y comercialización de envases para diferentes sectores industriales. Durante el año 2023, la organización había enfrentado un entorno empresarial cada vez más exigente, en el que la digitalización de procesos administrativos y financieros era indispensable para garantizar eficiencia, transparencia y competitividad. A pesar de contar con un sistema de facturación electrónica obligatorio por normativas nacionales, la empresa no lograba aprovechar de manera efectiva la información generada a partir de sus comprobantes electrónicos. Esto se debía a una serie de limitaciones internas y externas que afectaban la gestión administrativa, el control de operaciones y la capacidad de análisis estratégico.

- **Gestión Fragmentada de la Información**: Los procesos de facturación electrónica se encontraban poco integrados con otras áreas clave de la empresa, como contabilidad, logística y ventas. Esto generaba dificultades en la consolidación de información y en la elaboración de reportes financieros y comerciales que apoyaran la toma de decisiones oportunas.
- **Procesos Manuales Complementarios**: A pesar de la existencia de sistemas digitales, se mantenían tareas manuales de verificación y registro paralelo en hojas de cálculo. Esta duplicidad no solo incrementaba el margen de error humano, sino que también ocasionaba demoras en los procesos administrativos y un mayor costo operativo.

- **Acceso Limitado a Herramientas de Análisis de Datos**: La empresa carecía de sistemas especializados en minería de datos o inteligencia de negocios que permitieran transformar la información de las facturas en indicadores útiles. Como resultado, los registros electrónicos eran empleados únicamente para cumplir con obligaciones tributarias, desaprovechando su potencial para el análisis de clientes, productos y tendencias de mercado.
- **Deficiente Control de Clientes y Productos**: La falta de procesamiento avanzado de la información hacía que la empresa no pudiera segmentar a sus clientes de acuerdo con su nivel de compra o comportamiento de pago. Asimismo, no se lograba identificar con precisión qué productos presentaban mayor o menor rotación, lo que afectaba la planificación de ventas y la gestión del inventario.
- **Retrasos en la Detección de Irregularidades**: La ausencia de algoritmos predictivos dificultaba la identificación temprana de inconsistencias en las facturas, como anulaciones frecuentes, variaciones inusuales de montos o errores en la emisión. Esto ocasionaba contingencias tributarias y limitaba la capacidad de control interno.
- **Resistencia y Limitada Capacitación del Personal**: Parte del personal administrativo mostraba resistencia al uso de herramientas digitales avanzadas y carecía de formación en análisis de datos. Esta situación ralentizaba la adopción de nuevas tecnologías y limitaba la implementación de soluciones innovadoras como el clustering K-Means.
- **Oportunidades de Mejora Desaprovechadas**: El volumen significativo de datos generados por las facturas electrónicas representaba una fuente de información valiosa que no estaba siendo explotada. La carencia de técnicas de minería de datos impedía identificar oportunidades de optimización en políticas de precios, estrategias de fidelización de clientes y posicionamiento en el mercado regional.

1.2. Formulación del Problema

¿Cómo influirá la aplicación de la minería de datos mediante el algoritmo de clustering K-Means en la mejora de la gestión administrativa y la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023?

1.3. Objetivos

1.3.1. Objetivo General

Determinar la influencia de la minería de datos mediante el algoritmo de clustering K-Means en la mejora de la gestión administrativa y en la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023.

1.3.2. Objetivos Específicos

- Evaluar la calidad de los datos utilizados en el proceso de facturación electrónica.
- Evaluar la precisión de la clasificación realizada por el modelo K-Means, mediante el uso de métricas de validación.
- Determinar el número óptimo de clústeres generados por el algoritmo K-Means.
- Disminuir el tiempo promedio en la clasificación de facturas electrónicas.
- Cuantificar la exactitud en los reportes generados por el sistema de clasificación basado en clustering.
- Aumentar la eficiencia del modelo K-Means en la detección de patrones de facturación
- Determinar la tasa de mejora en la toma de decisiones contables.

1.4. Hipótesis

La aplicación de la minería de datos mediante el algoritmo de clustering K-Means influirá de manera significativa en la mejora de la gestión administrativa y en la optimización de la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023

1.5. Justificación

1.5.1. Justificación Teórica

a investigación se fundamenta en la necesidad de ampliar el conocimiento sobre la aplicación de técnicas de minería de datos en el ámbito empresarial, particularmente en el análisis de facturación electrónica mediante el algoritmo de clustering K-Means. Aunque existen estudios internacionales que demuestran la utilidad de la inteligencia artificial y el aprendizaje automático en la gestión de datos financieros y administrativos, en el contexto peruano se evidencia una brecha en el aprovechamiento de estas metodologías dentro de organizaciones medianas. La presente investigación aporta al marco teórico de la Ingeniería de Sistemas al integrar conceptos de segmentación de datos no supervisados con procesos administrativos reales, contribuyendo con evidencia empírica que refuerza y complementa la literatura existente sobre inteligencia de negocios y análisis de patrones en entornos empresariales.

1.5.2. Justificación Práctica

La relevancia de la investigación radica en que responde a una problemática concreta de la empresa Envases Los Pinos S.A.C., la cual genera un volumen considerable de información a través de su sistema de facturación electrónica, pero no explota dicho recurso de manera estratégica. La implementación de la minería de datos mediante clustering K-Means permite descubrir patrones de comportamiento en clientes y productos, optimizar procesos de gestión administrativa y facilitar decisiones estratégicas en áreas como ventas, logística e inventario. Este aporte resulta significativo no solo para la empresa objeto de estudio, sino también para otras organizaciones de la región y del país que enfrentan desafíos similares en la administración de grandes volúmenes de datos. En este sentido, la investigación genera un impacto práctico al demostrar cómo la facturación electrónica, más allá de su función tributaria, puede transformarse en una fuente clave de inteligencia empresarial.

1.5.3. Justificación Metodológica

La importancia de la investigación se sustenta en la aplicación de un enfoque cuantitativo que utiliza datos reales y relevantes de la organización, garantizando validez y aplicabilidad de los resultados. La elección del algoritmo de clustering K-Means se justifica por su eficacia en la segmentación de datos no supervisados y su adaptabilidad a contextos donde la información presenta heterogeneidad. Metodológicamente, la investigación ofrece una propuesta replicable y adaptable a otros entornos empresariales, lo que constituye un aporte significativo al campo de la Ingeniería de Sistemas. Al establecer un modelo que vincula técnicas de minería de datos con procesos administrativos, se fortalece la base metodológica disponible para futuros estudios y se contribuye a la consolidación del análisis de datos como herramienta estratégica en la gestión organizacional.

1.6. Importancia

La investigación reviste una importancia significativa porque aborda una problemática actual y relevante en el contexto empresarial, vinculada al aprovechamiento de la información contenida en las facturas electrónicas como un recurso estratégico para la toma de decisiones. En un entorno donde la digitalización y la transformación tecnológica se convierten en factores determinantes de competitividad, resulta imprescindible que las organizaciones adopten herramientas avanzadas de análisis de datos que les permitan gestionar de manera eficiente los grandes volúmenes de información que generan. En este sentido, el uso de la minería de datos mediante el algoritmo de clustering K-Means se presenta como una alternativa metodológica de gran valor, al posibilitar la identificación de patrones y la segmentación de clientes y productos con base en comportamientos reales reflejados en la facturación electrónica.

Desde una perspectiva académica y científica, la investigación es importante porque contribuye al fortalecimiento del conocimiento en el campo de la Ingeniería de Sistemas, integrando conceptos de inteligencia de negocios, minería de datos y aprendizaje automático con problemas reales de gestión empresarial. Al centrarse en un caso aplicado, como el de la empresa Envases Los Pinos S.A.C., se genera

evidencia empírica que enriquece la literatura existente y aporta un modelo replicable para estudios posteriores. Asimismo, la investigación promueve el desarrollo de enfoques interdisciplinarios, al vincular la gestión administrativa con la ciencia de datos, lo que fortalece el papel de la Ingeniería de Sistemas como disciplina clave en la transformación digital de las organizaciones.

En el plano práctico, la importancia del estudio radica en que proporciona soluciones concretas a una empresa que enfrenta dificultades en la explotación estratégica de su información. La aplicación del algoritmo de clustering K-Means permitirá segmentar clientes y productos, detectar tendencias de consumo y optimizar el control de operaciones, lo que se traducirá en una mejora sustancial en la gestión administrativa y en la planificación estratégica. Este aporte no solo impacta directamente en la eficiencia y sostenibilidad de la empresa en estudio, sino que también abre la posibilidad de que otras organizaciones de la región y del país implementen metodologías similares para aprovechar sus sistemas de facturación electrónica.

En términos sociales y económicos, la investigación adquiere relevancia porque contribuye al fortalecimiento de las capacidades tecnológicas en el ámbito empresarial peruano, promoviendo el uso de herramientas de análisis de datos que mejoran la transparencia, reducen errores administrativos y permiten decisiones más informadas. Esto se alinea con los esfuerzos nacionales de transformación digital y competitividad, generando beneficios no solo para las empresas, sino también para sus clientes, proveedores y la economía local en general.

Finalmente, la investigación es importante porque plantea una propuesta innovadora que trasciende la función tributaria de la facturación electrónica y la convierte en una fuente de inteligencia empresarial. Este enfoque aporta una nueva perspectiva sobre cómo los datos, al ser gestionados de manera adecuada mediante técnicas avanzadas, se transforman en un activo estratégico capaz de generar valor y de marcar la diferencia en la gestión organizacional en un entorno global altamente competitivo.

II. MARCO TEÓRICO

2.1. Antecedentes

2.1.1. Antecedentes Internacionales

Amari et al. (2024), tuvo como objetivo primordial diseñar un enfoque de deep learning que automatizara la validación de documentos de facturas electrónicas, reduciendo errores humanos y optimizando el proceso administrativo en entornos empresariales. La metodología consistió en aplicar modelos de redes neuronales convolucionales (CNN) y mecanismos de atención, entrenados con un conjunto de más de 50.000 facturas digitales en múltiples formatos. Los resultados evidenciaron que el sistema alcanzó una precisión del 94,6 % en la validación de facturas, mejorando en un 22 % el rendimiento frente a métodos tradicionales basados en reglas. Asimismo, se redujo en un 35 % el tiempo promedio de procesamiento por documento. Se concluyó que el uso de modelos profundos no solo mejora la eficiencia de la validación automática, sino que también fortalece la confiabilidad y escalabilidad de los sistemas de gestión documental en empresas con alto volumen de transacciones.

Tian et al. (2024), tuvieron como objetivo principal aplicar algoritmos de machine learning en la detección de fraudes fiscales y de facturación, específicamente en escenarios de evasión y manipulación de comprobantes. La investigación implementó modelos de clasificación supervisada, incluyendo Random Forest, Gradient Boosting y redes neuronales, evaluados con un conjunto de datos de transacciones fiscales y facturas electrónicas. Los resultados demostraron que los modelos lograron detectar actividades fraudulentas con una precisión del 93 %, lo que significó una mejora del 28 % respecto a los sistemas tradicionales de auditoría manual. Se concluyó que la integración de machine learning en la detección de fraudes fiscales representa una herramienta de alto valor para organismos de control, al reducir costos de supervisión y aumentar la confiabilidad de los procesos de fiscalización.

Arslan, Işık y Görmez (2024), tuvieron como objetivo principal desarrollar una solución de deep learning para la digitalización de imágenes de facturas y la generación automática de facturas electrónicas etiquetadas. La metodología consistió en implementar redes neuronales convolucionales y modelos de detección de objetos entrenados con un corpus de imágenes de facturas en múltiples idiomas y estructuras. Los resultados evidenciaron que el modelo alcanzó una tasa de reconocimiento del 91 % en campos relevantes y logró reducir en un 30 % el tiempo de procesamiento frente a sistemas OCR convencionales. Además, el sistema permitió generar facturas electrónicas estructuradas con una precisión del 88 %, contribuyendo a la automatización integral del ciclo documental. Se concluyó que este enfoque basado en deep learning ofrece una solución escalable y adaptable a distintos contextos empresariales, representando un avance significativo en la automatización de la facturación electrónica y la gestión inteligente de documentos.

Krieger, Drews y Funk (2023), tuvieron como objetivo primordial proponer un sistema de procesamiento automatizado de facturas que empleara algoritmos de aprendizaje automático para mejorar la extracción de información de proveedores de larga cola, caracterizados por facturas heterogéneas y no estandarizadas. La metodología consistió en un diseño experimental con facturas digitales de diferentes formatos y estructuras, aplicando técnicas de natural language processing (NLP), redes neuronales y gradient boosting. Los resultados indicaron que el sistema alcanzó una precisión del 93 % en la extracción de campos clave como montos, fechas y conceptos, lo que significó un incremento del 27 % frente a los métodos OCR tradicionales. Se concluyó que los modelos de machine learning aplicados a escenarios complejos de facturación permiten mejorar la trazabilidad de los procesos, reducir costos operativos y minimizar errores en la integración de datos financieros.

Schulte, Kieckbusch, Rocha Filho y Weigang (2022), desarrollaron un sistema denominado *ELINAC* que tuvo como objetivo primordial la agrupación eficiente de facturas electrónicas mediante autoencoders y técnicas de clustering. La metodología incluyó la aplicación de modelos de aprendizaje no supervisado sobre descripciones cortas de productos en facturas brasileñas, buscando reducir redundancias y mejorar la coherencia semántica. Los resultados mostraron que ELINAC mejoró en un 25 % la coherencia de los clústeres frente a métodos heurísticos tradicionales, además de optimizar en un 40 % el tiempo de procesamiento. Se concluyó que la combinación de autoencoders y clustering constituye un enfoque altamente eficiente para organizar grandes volúmenes de facturas electrónicas y detectar anomalías, aportando una herramienta estratégica para el análisis financiero y la toma de decisiones.

Bardelli, Rondinelli, Vecchio y Figini (2020), tuvieron como objetivo principal clasificar de manera automática facturas electrónicas empleando modelos de machine learning, con el propósito de optimizar la organización contable y minimizar la intervención manual. La metodología aplicada fue experimental, utilizando un conjunto de datos compuesto por miles de facturas electrónicas reales de empresas italianas, procesadas con algoritmos de support vector machines (SVM), random forest y redes neuronales. Los resultados demostraron que el modelo basado en random forest logró un desempeño superior, alcanzando una precisión del 89 %, lo que representó una mejora del 20 % respecto a los sistemas tradicionales. Se concluyó que la implementación de modelos de aprendizaje automático en la clasificación de facturas electrónicas constituye una estrategia eficaz para incrementar la productividad, reducir errores humanos y fortalecer los procesos de digitalización empresarial.

2.1.2. Antecedentes Nacionales

Torres Segovia (2024), tuvo como propósito desarrollar un sistema web basado en un modelo de recomendación con machine learning para apoyar el proceso de ventas en una ferretería de Chiclayo. El estudio utilizó un enfoque cuantitativo, de tipo aplicado y diseño experimental, implementando un sistema que integró algoritmos de filtrado colaborativo y análisis de patrones de compra en los clientes. Los resultados indicaron que las recomendaciones generadas aumentaron en un 28 % el promedio de ventas mensuales y mejoraron en un 31 % la satisfacción de los clientes encuestados. Se concluyó que el uso de sistemas de recomendación basados en machine learning fortalece la relación con los clientes y aporta a la optimización de procesos de comercialización en pequeñas y medianas empresas, consolidando la importancia de la inteligencia artificial como apoyo estratégico para la competitividad en mercados locales.

Suárez Romero (2024), se propuso diseñar un modelo de Deep Learning orientado a mejorar la predicción de ventas en la empresa San Fernando S.A.C. en Lima durante el año 2023. La metodología aplicada fue cuantitativa y de diseño experimental, utilizando redes neuronales recurrentes (RNN) y bases de datos históricas de ventas para entrenar el modelo. El análisis comparativo con modelos tradicionales de regresión mostró que el Deep Learning alcanzó una precisión del 94 % en la predicción, lo que significó una mejora del 22 % respecto a los modelos convencionales utilizados previamente por la empresa. Se concluyó que la implementación de modelos de Deep Learning no solo mejora la capacidad predictiva en escenarios de alta variabilidad de la demanda, sino que también constituye un soporte fundamental para la planificación estratégica y la toma de decisiones en grandes corporaciones del sector alimenticio.

Falén Ordinola, Aquino Trujillo y Castillo Montalván (2024), tuvieron como objetivo central aplicar modelos de machine learning para optimizar el proceso de facturación en una empresa de servicios. La investigación se sustentó en un enfoque aplicado y de diseño experimental, donde se implementó un modelo de clasificación supervisada para identificar

patrones en los errores de facturación. Los resultados demostraron una reducción del 37 % en las inconsistencias detectadas y una mejora del 29 % en la eficiencia de los tiempos de facturación. Se concluyó que la incorporación de machine learning en los procesos de facturación constituye una estrategia viable para mejorar la exactitud de los registros, disminuir costos operativos y optimizar la gestión administrativa, lo que confirma el valor de la inteligencia artificial en los procesos contables y financieros en el ámbito empresarial peruano.

Ávila Galindo (2023), tuvo como objetivo principal determinar la influencia de la Robotic Process Automation (RPA) en la productividad del área de pagos en una empresa del sector retail en Lima. La investigación siguió una metodología aplicada, con un diseño experimental, aplicando un sistema de automatización de procesos en la gestión de pagos a proveedores. Para la recolección de datos se compararon indicadores de tiempo de procesamiento y número de errores antes y después de la implementación del RPA. Los resultados mostraron que la productividad del área se incrementó en un 35 %, mientras que los errores humanos en el procesamiento de pagos se redujeron en un 42 %. Se concluyó que la implementación de RPA representa una herramienta eficaz para optimizar procesos administrativos, permitiendo liberar recursos humanos de tareas repetitivas y aumentando la eficiencia en la gestión financiera, lo cual constituye un precedente relevante para la integración de nuevas tecnologías en áreas administrativas de las empresas peruanas.

2.1.3. Antecedentes Locales

No existen investigaciones locales en repositorios académicos.

2.2. Marco Conceptual

2.2.1. Teoría de Conglomerados

Kaufman & Rousseeuw (1990), definen el análisis de conglomerados como "el conjunto de métodos estadísticos destinados a descubrir estructuras de agrupamiento en los datos, dividiendo un conjunto de observaciones en grupos (clusters) que son internamente homogéneos y externamente heterogéneos".

Por otro lado, Han, Kamber, & Pei (2011) explican que el análisis de clúster es "una técnica de minería de datos que organiza un conjunto de objetos en grupos, de modo que los objetos en el mismo grupo sean más similares entre sí que con los de otros grupos"; además, destaca su papel como herramienta exploratoria para identificar estructuras naturales.

En síntesis, el análisis de conglomerados es una técnica estadística multivariante que agrupa elementos con características similares en grupos o clusters, de modo que los objetos dentro de un cluster sean homogéneos y heterogéneos respecto a otros clusters. Es una técnica exploratoria y descriptiva ampliamente usada para segmentar datos en diferentes campos, desde la administración hasta la biología. (Ruiz Aranibar, 2019).

2.2.2. Minería de Datos

Han, Kamber, & Pei (2011) destacan que la minería de datos comprende una variedad de técnicas entre ellas el análisis de clúster cuyo objetivo es "descubrir patrones interesantes, desconocidos y útiles a partir de grandes volúmenes de datos". Asimismo, Han & Kamber (2006), enfatizan que el cluster analysis es una técnica central de minería de datos descriptiva y exploratoria, utilizada para organizar datos sin conocimiento previo y descubrir estructuras útiles como base para modelos predictivos.

2.2.3. Machine Learning

Grupo de métodos computacionales que utilizan la experiencia (información pasada disponible) hacer predicciones o mejorar el rendimiento. La calidad y tamaño de los datos son cruciales para el éxito de las predicciones realizadas. Consiste en diseñar algoritmos de predicción eficientes y precisos. Algunas medidas críticas de la calidad de estos

algoritmos son su complejidad de espacio y tiempo. Además, se necesita una noción de complejidad de la muestra para evaluar su tamaño y que el algoritmo aprenda una familia de conceptos. Generalmente, las garantías de aprendizaje teórico para un algoritmo dependen de la complejidad de las clases de concepto consideradas y del tamaño de la muestra de entrenamiento. Machine learning está inherentemente relacionado con el análisis de datos y las estadísticas dado que el éxito de un algoritmo de aprendizaje depende de los datos utilizados (Mohri et al., 2018)

2.2.3.1. Categorías

Los algoritmos de machine learning se pueden categorizar en función de si necesitan etiquetas (presencia en los datos de un resultado ideal que un modelo debería producir para un ejemplo dado), según esto pueden ser (Mohri et al., 2018):

a) **Supervisados**. Aprovechan los conjuntos de datos que contienen etiquetas para las entradas y su objetivo es aprender un mapeo de las entradas a las etiquetas. +

   Está diseñado para encontrar patrones en datos etiquetados. Debido a que se han identificado los atributos y el significado de los datos, los usuarios que están entrenando los datos modelados lo entienden bien para que se ajusten a los detalles de las etiquetas. Cuando la etiqueta es continua, es una regresión; cuando los datos provienen de un conjunto finito de valores, se conoce como clasificación. En esencia, la regresión utilizada para el aprendizaje supervisado lo ayuda a comprender la correlación entre las variables. Los algoritmos se entrenan usando ejemplos preprocesados. El rendimiento del algoritmo se evalúa con datos de prueba. A veces los patrones que se identifican en un subconjunto de datos no se pueden detectar en la población de datos más grande. Cuando el modelo se ajusta con precisión a los datos de entrenamiento para representar solo los patrones que existen en el subconjunto puede que no llegue a ser aplicable a grandes conjuntos de datos desconocidos, A esto se le conoce como sobreajuste. Para

protegerse contra el sobreajuste, las pruebas deben realizarse con datos etiquetados imprevistos o desconocidos. El uso de datos imprevistos para el conjunto de prueba ayuda a evaluar la precisión del modelo en la predicción de resultados y conclusiones. Los modelos de capacitación supervisada tienen una amplia aplicabilidad a una variedad de problemas comerciales (Hurwitz & Kirsch, 2018).

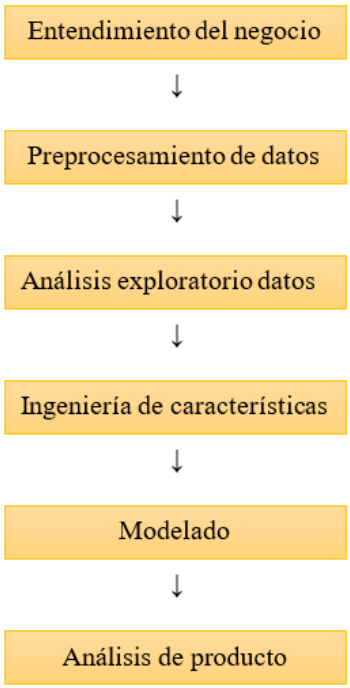
b) No supervisados. No requieren etiquetas.

Su uso es más adecuado cuando el problema requiere una gran cantidad de datos que no están etiquetados. Entender el significado detrás de los datos requiere que los algoritmos puedan comprender el significado en función de poder clasificar los datos según los patrones que encuentre. Segmentan los datos en grupos de características denominados clusters. Los datos sin etiquetar crean los valores de los parámetros y la clasificación de los datos. Es decir que este proceso agrega etiquetas a los datos para que puedan ser supervisados. Puede determinar el resultado cuando hay una gran cantidad de datos. Como el desarrollador no conoce el contexto de los datos que se van a analizar el etiquetado no es posible en esta etapa. Es por eso que el aprendizaje no supervisado puede ser utilizado antes de pasar los datos a un proceso de aprendizaje supervisado. Pueden ayudar a las empresas a comprender grandes volúmenes de datos sin etiquetar. A diferencia del aprendizaje supervisado estos algoritmos buscan patrones en datos que aún no se entienden. Es por eso que un enfoque de aprendizaje no supervisado puede ayudar a determinar los resultados más rápidamente que un enfoque de aprendizaje supervisado. Para implementar un proyecto de Machine Learning no supervisado se deben llevar a cabo seis etapas consecutivas (Hurwitz & Kirsch, 2018):

- Entendimiento del negocio.
- Preprocesamiento de datos.
- Análisis exploratorio de datos
- Ingeniería de características.
- Modelado.
- Análisis de producto.

Figura 2:

Etapas de implementación de Machine Learning



2.2.4. Algoritmo K-Means

El algoritmo K-Means es una técnica fundamental dentro del campo de la minería de datos, específicamente categorizada como un método de aprendizaje no supervisado, cuyo propósito principal es agrupar un conjunto de observaciones en un número determinado de clústeres (o agrupaciones) distintos, de manera que cada observación pertenezca al clúster con la media más cercana, considerada como su prototipo (Bishop, 2006).

2.2.5. Facturación Electrónica

Documento digital que cumple los propósitos de una factura en papel y que registra las operaciones comerciales de una entidad en manera electrónica cumpliendo, en cualquier situación que aplique y ante todos los actores, los principios de autenticidad, integridad y legibilidad del proceso. Al ser electrónica tiene algunas características y condiciones propias como (Barreix & Zambrano, 2018):

- Es almacenado y transmitido a través de medios electrónicos.
- No existen diferencias entre el original y la copia reproducida.
- Se interpretar la estructura a través de reglas y procesos definidos.

2.2.6. Sistema de emisión de comprobantes de pago electrónicos

Iniciativa dada por SUNAT que reemplaza los documentos emitidos en papel por una integración electrónica entre los procesos de los contribuyentes y los de SUNAT. Este sistema promueve beneficios para los negocios y trabajadores independientes tales como:

- Ahorro en costos.
- Optimización de recursos.
- Mejora de procesos.

Está formado por una multiplataforma de emisión electrónica que puede ser utilizada por todo tipo de contribuyente y acorde a sus operaciones comerciales y de facturación, sean obligados o voluntarios. A través de la multiplataforma, el contribuyente puede emitir facturas, boletas de venta, notas de crédito y débito, guías de remisión, recibos por servicios públicos, comprobantes de retención del IGV y comprobantes de percepción del IGV. No son sistemas excluyentes por lo que el contribuyente puede utilizar uno o más de uno (de Velazco Borda, 2016).

2.2.6.1. Sistemas de emisión electrónica - SUNAT operaciones en línea

Aplicativo de la web de SUNAT en el cual se pueden emitir gratuitamente comprobantes de pago electrónicos. Para este aplicativo es indispensable tener una Clave SOL y registrar la información solicitada. Está pensado para pequeñas y microempresas (PYME) que emiten un número reducido de

comprobantes y para trabajadores independientes que emiten recibos por honorarios. La finalidad es facilitar el cumplimiento de las obligaciones tributarias vinculadas a la emisión de comprobantes de pago, registro de libros contables y declaración los impuestos. Las principales características con las que cuenta este sistema son (de Velazco Borda, 2016):

- Gratuito.
- Accesible por internet.
- Serie y correlativo son generados automáticamente.
- No hay obligación de conservar los comprobantes.
- Consulta y descarga del comprobante en cualquier momento.
- Envió de copia por correo electrónico.

2.2.6.2. Sistemas de emisión electrónica - Sistemas del Contribuyente

Dirigido a los negocios que emiten gran número de comprobantes de pago, les permite emitir desde sistemas propios o tercerizados. En este sistema son los contribuyentes los que deben elaborar los comprobantes de pago electrónicos de acuerdo a los requisitos, características y condiciones definidos por SUNAT, para luego ser enviados y validados en línea a fin de darles la validez tributaria correspondiente. Sus principales características son (de Velazco Borda, 2016).:

- Debe ser hecho en un archivo XML, basado en el estándar UBL versión 2.0.
- Debe ser firmado digitalmente con un certificado digital.
- Debe ser enviados a través de un web service.
- La serie y el correlativo son definidos por el emisor.
- El emisor y receptor están obligados a conservar el comprobante.

2.2.6.3. Otras aplicaciones

Desde la APP SUNAT se pueden emitir comprobantes de pago electrónicos desde cualquier dispositivo Android o IOS. Además, se ha puesto a disposición del contribuyente el llamado Facturador

SUNAT, una nueva aplicación gratuita que está dirigida a medianos y pequeños negocios y cuya principal característica es que puede integrarse al sistema de la empresa para generar y enviar de manera automática de comprobantes de pago. El contribuyente sólo requiere habilitar la captura de los datos de la transacción, y la aplicación genera el formato XML según la estructura requerida y también realiza el proceso de firma digital, para lo cual se debe contar previamente con un certificado digital. La generación del formato XML puede realizarse sin conexión a internet. Dado que el sistema cuenta con todas las reglas de validaciones necesarias la probabilidad de error es nula. Por su parte, pensando en el receptor se ha implementado una consulta a través del web service para que verificación de validez de comprobantes de pago electrónicos (de Velazco Borda, 2016).

2.2.7. Segmentación de clientes

Es la división de la base de clientes en grupos llamados segmentos de manera que cada uno consiste en clientes que comparten características similares de mercado. Estas distinciones se basan en factores que influyen directa o indirectamente en el negocio como preferencias de productos, ubicaciones, comportamiento, etc. La importancia de la segmentación de clientes incluye (Vijilesh y otros, 2021):

- La capacidad de una empresa para personalizar los planes de mercado que serán apropiados para cada segmento de sus clientes.
- Apoyo a las decisiones empresariales basadas en un entorno de riesgo como las relaciones de endeudamiento con sus clientes.
- Identificación de productos relacionados con componentes individuales y cómo administrar la demanda y el suministro de energía.
- revela la interdependencia y la interacción entre consumidores, entre productos o entre clientes y productos de los que la empresa puede no estar al tanto.
- La capacidad de predecir la disminución de clientes y qué clientes tienen más probabilidades de tener problemas y plantear otras preguntas de

investigación de mercado y proporcionar pistas para encontrar soluciones.

La segmentación de clientes ayuda a una organización de distintas maneras. A continuación, se enumeran los principales objetivos y beneficios detrás de la motivación para la segmentación de clientes.

2.2.7.1. Comprensión del cliente

La comprensión más profunda de los clientes de una empresa, sus atributos y comportamiento es uno de los objetivos principales de un proceso de segmentación de clientes. Esta información sobre la base de clientes se puede utilizar de distintas formas y por si misma ya es útil. Uno de los paradigmas empresariales más aceptados es el de "conoce a tu cliente" y una segmentación de la base de clientes permite una disección perfecta de este paradigma. Esta comprensión y su explotación es lo que forma la base de los otros beneficios de la segmentación de clientes (Sarkar et al., 2018).

2.2.7.2. Marketing objetivo

La capacidad de enfocar los esfuerzos de marketing de manera eficaz y eficiente es la razón más visible para la segmentación de clientes. Para diseñar mejores campañas de marketing se tiene que conocer los diferentes segmentos de su base de clientes. Un buen modelo de segmentación permite una mejor comprensión de los requisitos del cliente y, por lo tanto, aumenta las posibilidades de éxito de cualquier campaña de marketing (Sarkar et al., 2018).

2.2.7.3. Colocación óptima de productos

Una buena estrategia de segmentación de clientes también puede ayudar a la empresa a desarrollar u ofrecer nuevos productos. Este beneficio depende en gran medida de la forma en que se aproveche el proceso de segmentación (Sarkar et al., 2018).

2.2.7.4. Búsqueda de segmentos de clientes latentes

Un proceso de segmentación de clientes ayuda a la organización a conocer su base de clientes. Un efecto secundario obvio de cualquier práctica de este tipo es descubrir qué segmento de clientes podría estar faltando. Esto puede ayudar a identificar segmentos de clientes sin explotar centrándose en campañas de marketing o desarrollo de nuevos productos (Sarkar et al., 2018).

2.2.7.5. Mayores ingresos

Este es el requisito más importante. El motivo es que la segmentación de clientes puede generar mayores ingresos debido a los efectos combinados de todos los objetivos anteriores (Sarkar et al., 2018).

2.2.8. Metodología CRISP-DM

El proceso de minería de datos, especialmente cuando se aplica a la optimización de la gestión administrativa y la toma de decisiones en entornos empresariales complejos, requiere de un marco metodológico estructurado que guíe las diversas etapas del proyecto desde su concepción hasta su implementación y evaluación. En este contexto, la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) emerge como un estándar ampliamente reconocido y adoptado en la industria y la academia, proporcionando un enfoque sistemático y flexible para la realización de proyectos de minería de datos (Siddiqui et al., 2020)

2.2.8.1. Fases de la metodología CRISP-DM

CRISP-DM se caracteriza por su naturaleza iterativa y cíclica, lo que permite a los equipos de proyecto refinar continuamente sus enfoques y adaptarse a las nuevas comprensiones que surgen a medida que avanza el análisis de los datos. La metodología se divide en seis fases principales:

- Comprensión del Negocio (Business Understanding).
- Comprensión de los Datos (Data Understanding).
- Preparación de los Datos (Data Preparation).
- Modelado (Modeling).
- Evaluación (Evaluation).
- Despliegue (Deployment).

Cada una de estas fases es crucial para el éxito del proyecto, y la interconexión entre ellas permite abordar los desafíos de manera integral. Por ejemplo, la profunda comprensión del negocio y de los datos es fundamental para definir los objetivos del modelado y diseñar estrategias efectivas de preparación de datos.

III. METODOLOGÍA

3.1. Enfoque

El enfoque de la investigación es cuantitativo, ya que se fundamenta en la recolección y análisis de datos numéricos provenientes de los registros de facturación electrónica de la empresa Envases Los Pinos S.A.C. El propósito es identificar patrones, tendencias y agrupamientos mediante la aplicación del algoritmo de clustering K-Means, lo que permite obtener resultados medibles y verificables. De acuerdo con Huamán (2022), el enfoque cuantitativo se caracteriza por estructurar los datos en indicadores objetivos que pueden ser analizados estadísticamente, garantizando precisión y confiabilidad en la interpretación de los hallazgos. En ese sentido, este enfoque se adecúa a investigaciones en ingeniería de sistemas que buscan optimizar procesos administrativos mediante técnicas de minería de datos.

3.2. Método

El método que guía el estudio es hipotético-deductivo, dado que parte de la formulación de una hipótesis sobre la influencia del algoritmo de clustering K-Means en la gestión administrativa y la toma de decisiones estratégicas. Posteriormente, se procede a la recolección y análisis de datos de facturación para contrastar empíricamente la hipótesis planteada. Según Ramos (2021), este método se aplica en proyectos que requieren comprobar relaciones de causalidad entre variables, lo que resulta pertinente para evaluar el impacto de técnicas de machine learning en entornos empresariales. A través de este método, se garantiza un proceso lógico que avanza desde la teoría hasta la verificación práctica.

3.3. Diseño

El tipo de investigación es aplicada de nivel explicativo. Es aplicada porque busca resolver un problema real en la empresa Envases Los Pinos S.A.C., optimizando la gestión administrativa mediante el uso de algoritmos de machine learning. A su vez, es explicativa porque pretende establecer la relación de causa-efecto entre la implementación del algoritmo de clustering K-Means (variable independiente) y la mejora de la gestión administrativa y toma de decisiones (variable dependiente). Según Mamani (2023), este tipo de investigación permite no solo describir un

fenómeno, sino también comprender los factores que lo determinan y evaluar la magnitud de su influencia. En este caso, se espera que los resultados ofrezcan evidencias que respalden la adopción de tecnologías de minería de datos como herramientas estratégicas de gestión.

3.4. Población

La población de estudio estuvo conformada por la totalidad de facturas electrónicas emitidas por la empresa Envases Los Pinos S.A.C. durante el año 2023. De acuerdo con los registros de su sistema de contabilidad electrónica, esta población ascendió aproximadamente a 12,500 facturas electrónicas, que incluyen transacciones de ventas nacionales e internacionales, con diferentes montos, clientes y condiciones de pago.

3.5. Muestra

La muestra seleccionada estuvo constituida por 1,250 facturas electrónicas, lo que representó el 10 % del total de la población. Esta proporción permitió obtener un conjunto de datos manejable para el análisis, garantizando la representatividad de los distintos tipos de transacciones registradas en la empresa durante el periodo de estudio.

3.6. Muestreo

El muestreo aplicado fue no probabilístico por conveniencia, ya que las facturas fueron seleccionadas considerando su disponibilidad en el sistema y su relevancia para el análisis de patrones de facturación. Se optó por este tipo de muestreo debido a que el estudio se orientó a la aplicación del algoritmo K-Means sobre un conjunto significativo de datos, más que a la generalización estadística de los resultados.

3.7. Variables de Estudio

- Variable Dependiente:
- Variable Independiente:

3.8. Operacionalización de variables

Tabla 1:

Operacionalización de las Variables

Variable	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Escala de medición
Minería de datos mediante el algoritmo de clustering K-Means (Variable independiente)	Conjunto de técnicas que permiten explorar grandes volúmenes de datos, identificando patrones y agrupaciones similares sin supervisión, mediante el algoritmo K-Means (Tan, Steinbach & Kumar, 2019).	Aplicación del algoritmo de clustering K-Means sobre los datos de facturación electrónica de la empresa, generando grupos homogéneos que permitan identificar comportamientos de clientes, montos y frecuencia de compras.	Metodología CRISP - DM	Calidad de los datos	Razón
				Precisión en la clasificación	Razón
				Número de clúster generados	Razón
Gestión administrativa y toma de decisiones estratégicas basadas en la información de las facturas electrónicas (Variable dependiente)	Conjunto de procesos orientados a planificar, organizar y controlar la información contable y de facturación de la empresa, para apoyar la toma de decisiones estratégicas (Koontz & Weihrich, 2016).	Uso de los resultados del clustering para mejorar la gestión administrativa de las facturas electrónicas y apoyar la toma de decisiones estratégicas en la empresa Envases Los Pinos S.A.C.	Eficiencia Operativa	Tiempo en la clasificación de facturas electrónicas	Razón
			Confiabilidad	Exactitud en los reportes generados	Razón
			Capacidad de análisis	Eficiencia en la detección de patrones de facturación	Razón
			Impacto del modelo	Tasa de mejora en la toma de decisiones	Razón

3.9. Matriz de Consistencia

Tabla 2:

Matriz de consistencia

Problema General	Hipótesis General	Objetivo General	Objetivos Específicos	Técnicas e Instrumentos
¿Cómo influirá la aplicación de la minería de datos mediante el algoritmo de clustering K-Means en la mejora de la gestión administrativa y la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023?	La aplicación de la minería de datos mediante el algoritmo de clustering K-Means influirá de manera significativa en la mejora de la gestión administrativa y en la optimización de la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023	Determinar la influencia de la minería de datos mediante el algoritmo de clustering K-Means en la mejora de la gestión administrativa y en la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023	- Evaluar la calidad de los datos utilizados en el proceso de facturación electrónica. - Evaluar la precisión de la clasificación realizada por el modelo K-Means, mediante el uso de métricas de validación. - Determinar el número óptimo de clústeres generados por el algoritmo K-Means. - Disminuir el tiempo promedio en la clasificación de facturas electrónicas. - Cuantificar la exactitud en los reportes generados por el sistema de clasificación basado en clustering. - Aumentar la eficiencia del modelo K-Means en la detección de patrones de facturación - Determinar la tasa de mejora en la toma de decisiones contables.	- Análisis Documental - Cronometría - Registro - Lista de Errores - Validación Estadística

3.10. Técnicas e Instrumentos de recolección de datos

3.10.1. Técnicas de recolección de datos

- **Análisis Documental**: se utiliza para recopilar y organizar los datos provenientes del sistema contable y de facturación electrónica de la empresa. Esta técnica permite acceder a los registros históricos de transacciones emitidas durante el año 2023, garantizando que la información recolectada sea válida, confiable y representativa de la realidad administrativa.
- **Minería de datos**: posibilita la extracción, transformación y carga (ETL) de la información, con el objetivo de preparar los datos para su procesamiento mediante el algoritmo de clustering K-Means. Esta técnica es fundamental para identificar patrones ocultos y segmentar los registros de facturación en grupos homogéneos que faciliten la interpretación de comportamientos comerciales.

3.10.2. Instrumentos de recolección de datos

- **Base de datos Relacional**: Proporcionada por el sistema de facturación electrónica de la empresa, en la cual se encuentran registradas variables como fecha de emisión, cliente, monto, forma de pago y estado de la factura. Adicionalmente, se utiliza el software Python con librerías especializadas en aprendizaje automático (Scikit-learn, Pandas y Matplotlib) como instrumento analítico para la implementación del algoritmo K-Means y la validación de los clústeres mediante métricas
- **Cuestionario Estructurado**: dirigido a los responsables administrativos de la empresa, con el fin de evaluar el nivel de satisfacción y utilidad percibida respecto a los reportes generados a partir del análisis de clustering. Este cuestionario se diseña bajo escala tipo Likert de cinco niveles, lo que permite medir de manera cuantitativa la percepción de la gerencia en relación con la mejora de la gestión administrativa y la toma de decisiones estratégicas

3.11. Técnicas de Análisis de resultados

A. Estadística Descriptiva

A través de medidas como frecuencias, porcentajes, promedios y desviación estándar, con el propósito de caracterizar las facturas electrónicas según variables como montos, fechas, clientes y condiciones de pago. Esta técnica permite obtener una visión general del comportamiento de la información y su distribución.

B. Análisis de clustering mediante el algoritmo K-Means

Para segmentar las facturas electrónicas en grupos homogéneos. El desempeño del algoritmo se evalúa a través de métricas específicas como el Silhouette Score, que mide la cohesión y separación de los clústeres, y el Error Cuadrático Medio (SSE), que cuantifica la compacidad de cada grupo. Dichas métricas permiten determinar la calidad de los patrones generados y validar la pertinencia del número de clústeres seleccionados.

C. Visualización de datos

mediante gráficos de dispersión, diagramas de barras y representaciones multidimensionales, elaborados con el software Python y librerías como Matplotlib y Seaborn. Este recurso facilita la interpretación visual de los resultados obtenidos del clustering, posibilitando que la gerencia de la empresa identifique tendencias, clientes prioritarios o irregularidades en la facturación.

D. Análisis Comparativo

Se realiza entre los procesos administrativos previos y los resultados obtenidos con la aplicación del algoritmo, lo que permite medir mejoras en eficiencia, reducción de tiempos de clasificación y aumento de la calidad de la información utilizada para la toma de decisiones. Esta comparación se complementa con los resultados de los cuestionarios aplicados a los responsables administrativos, cuyos datos son procesados mediante análisis de frecuencias y escala Likert.

3.12. Consideraciones Éticas

La investigación se desarrolla bajo el cumplimiento estricto de principios éticos que garantizan la integridad científica, la protección de los datos y el respeto hacia la empresa objeto de estudio. En primer lugar, se asegura la confidencialidad y anonimato de la información contenida en las facturas electrónicas de la empresa Envases Los Pinos S.A.C. Para ello, los datos utilizados en el proceso de minería se codifican y procesan sin exponer información sensible de clientes, proveedores o montos específicos que puedan comprometer la privacidad corporativa o individual.

Asimismo, se observa el principio de consentimiento informado, mediante la autorización formal otorgada por la gerencia de la empresa para acceder a las bases de datos de facturación. Este consentimiento garantiza que la investigación se realiza con pleno conocimiento y conformidad de la organización, evitando el uso indebido de los datos.

De igual modo, se respeta el principio de transparencia y honestidad científica, evitando la manipulación de resultados y asegurando que las conclusiones correspondan al análisis riguroso de los datos obtenidos. El uso de herramientas tecnológicas, como algoritmos de clustering y software de análisis, se limita exclusivamente a fines académicos e investigativos, sin fines comerciales ni de lucro personal.

Finalmente, se cumple con las normativas nacionales e institucionales vigentes relacionadas con la protección de datos personales y el tratamiento de información digital, garantizando que los resultados obtenidos sean aplicables y útiles para la mejora de los procesos administrativos, sin vulnerar los derechos de la empresa ni de terceros involucrados.

IV. RESULTADOS Y DISCUSIÓN

4.1. Resultados

4.1.1. Metodología CRISP-DM

4.1.1.1. Comprensión del Negocio

A. Planteamiento del problema empresarial

En el entorno empresarial actual, caracterizado por una competencia creciente y un mercado cada vez más segmentado, las organizaciones generan y almacenan grandes volúmenes de información sobre las transacciones con sus clientes. Un ejemplo claro es la facturación electrónica, que, además de cumplir con los requisitos fiscales y legales, concentra datos valiosos sobre las compras realizadas.

No obstante, en la práctica, gran parte de esta información permanece sin analizar o se utiliza únicamente con fines administrativos y contables, dejando de lado su potencial para generar conocimiento estratégico. La falta de procesos y herramientas para transformar estos datos en información accionable limita la capacidad de las organizaciones para identificar patrones de comportamiento, segmentar a sus clientes y diseñar estrategias personalizadas de retención, fidelización y captación.

Esta situación se traduce en la aplicación de estrategias comerciales generalizadas que no consideran las diferencias en recencia de compra, frecuencia de consumo o monto gastado. Como resultado se desperdician recursos se pierde la oportunidad de fortalecer la relación con clientes de alto valor y se deja de actuar proactivamente frente a clientes inactivos o en riesgo de abandono.

Si bien existen técnicas de análisis de datos y aprendizaje automático como el algoritmo K-Means Clusteringque permiten segmentar clientes en grupos homogéneos basados en variables clave, su adopción en muchas organizaciones es baja debido al desconocimiento, la falta de capacitación o la

ausencia de integración de estas metodologías en los procesos de toma de decisiones. Esto genera una brecha entre el volumen de datos disponibles y la capacidad real de utilizarlos para obtener una ventaja competitiva sostenible.

El problema empresarial en Envases Los Pinos S.A.C. se centra en la complejidad de la gestión de su creciente volumen de facturas electrónicas, lo que dificulta identificar patrones de consumo, segmentar adecuadamente a los clientes y optimizar la toma de decisiones estratégicas. A pesar de cumplir con las normativas de facturación electrónica impuestas por la SUNAT, la empresa utiliza los comprobantes principalmente con fines tributarios, desaprovechando su valor potencial como fuente de inteligencia de negocio. Esta limitación ha generado una gestión fragmentada de la información, retrasos en la identificación de irregularidades y ausencia de segmentación que apoye estrategias comerciales y de fidelización.

B. Objetivos

El objetivo de este análisis es segmentar a los clientes en grupos homogéneos utilizando el algoritmo de clustering K-Means, con el fin de identificar patrones de comportamiento en sus compras. Esto permitirá desarrollar estrategias personalizadas para mejorar la retención, aumentar las ventas y optimizar la gestión de clientes.

C. Criterios de Éxito

Los criterios de éxito se establecieron en la capacidad del modelo para identificar un número óptimo de clústeres (k=4) con alta cohesión y separación (Silhouette Score de 0.642), lo que permite clasificar a los clientes en perfiles como inactivos, activos de bajo valor, clientes recurrentes de gasto medio y clientes VIP de alta frecuencia y alto gasto. El mapeo de estos resultados al negocio se traduce en estrategias diferenciadas: reactivación de clientes inactivos, programas de fidelización

para clientes activos de bajo gasto, y planes exclusivos para clientes VIP que concentran la mayor rentabilidad.

D. Supuestos y Limitaciones

Los supuestos que guían esta fase consideran que los datos de facturación son completos y consistentes, y que las métricas RFM reflejan de manera adecuada el comportamiento de los clientes. Sin embargo, existen limitaciones específicas como la resistencia del personal a adoptar nuevas herramientas analíticas, la posible falta de integración con sistemas contables y logísticos, y el riesgo de sesgo en la muestra de facturas seleccionadas. Estas restricciones condicionan la interpretación de los hallazgos y deben tenerse en cuenta al trasladar los resultados a la práctica empresarial.

4.1.1.2. Comprensión de los datos

A. Descripción de la fuente de datos

Los registros provinieron del sistema de facturación electrónica y contabilidad de la empresa, el cual almacena comprobantes de pago emitidos entre enero y diciembre de 2023. La base de datos contenía aproximadamente 12,500 facturas electrónicas, emitidas tanto a nivel nacional como internacional, y gestionadas a través de un sistema ERP interno con validación en línea por la SUNAT.

Se respetó la confidencialidad de los datos, utilizando únicamente la información necesaria para el análisis de los identificadores de clientes.

B. Recolección de datos

Se utilizó como fuente el archivo D_VENTAS_FACTURA_ENCA.xlsx, que contiene el registro histórico de facturas electrónicas emitidas por la empresa.

Figura 3:

Selección del Dataset

Figura 4:

Selección del Archivo

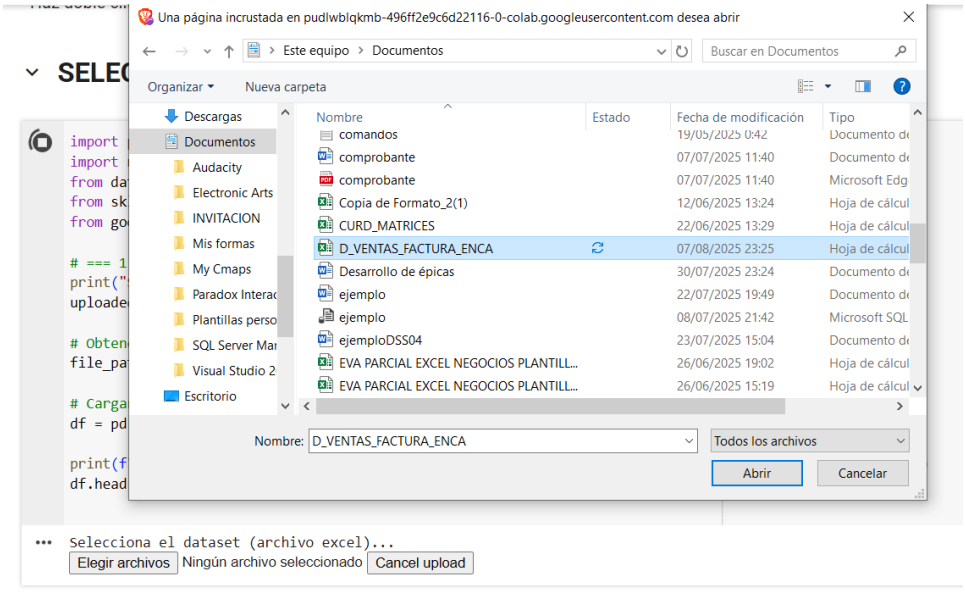
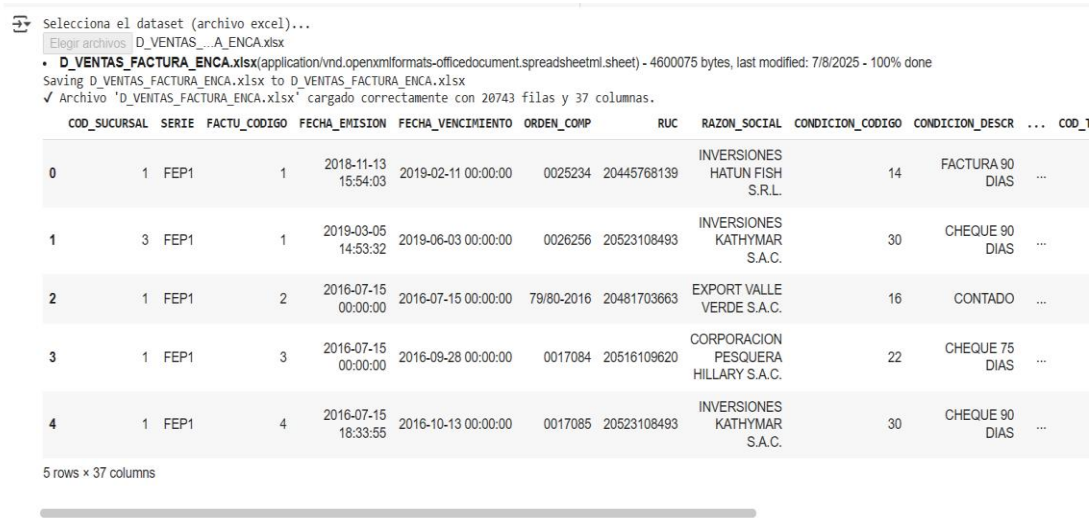


Figura 5:

Archivo Cargado



```
Selecciona el dataset (archivo excel)...
Elegir archivos D_VENTAS_...A_ENCA.xlsx
• D_VENTAS_FACTURA_ENCA.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 4600075 bytes, last modified: 7/8/2025 - 100% done
Saving D_VENTAS_FACTURA_ENCA.xlsx to D_VENTAS_FACTURA_ENCA.xlsx
✓ Archivo 'D_VENTAS_FACTURA_ENCA.xlsx' cargado correctamente con 20743 filas y 37 columnas.
```

	COD_SUCURSAL	SERIE	FACTU_CODIGO	FECHA_EMISION	FECHA_VENCIMIENTO	ORDEN_COMP	RUC	RAZON_SOCIAL	CONDICION_CODIGO	CONDICION_DESCR	...	COD_T
0	1	FEP1	1	2018-11-13 15:54:03	2019-02-11 00:00:00	0025234	20445768139	INVERSIONES HATUN FISH S.R.L.	14	FACTURA 90 DIAS	...	
1	3	FEP1	1	2019-03-05 14:53:32	2019-06-03 00:00:00	0026256	20523108493	INVERSIONES KATHYMAR S.A.C.	30	CHEQUE 90 DIAS	...	
2	1	FEP1	2	2016-07-15 00:00:00	2016-07-15 00:00:00	79/80-2016	20481703663	EXPORT VALLE VERDE S.A.C.	16	CONTADO	...	
3	1	FEP1	3	2016-07-15 00:00:00	2016-09-28 00:00:00	0017084	20516109620	CORPORACION PESQUERA HILLARY S.A.C.	22	CHEQUE 75 DIAS	...	
4	1	FEP1	4	2016-07-15 18:33:55	2016-10-13 00:00:00	0017085	20523108493	INVERSIONES KATHYMAR S.A.C.	30	CHEQUE 90 DIAS	...	

5 rows × 37 columns

C. Diccionario de datos

Los principales campos registrados fueron:

- número de factura
- fecha de emisión
- RUC del cliente
- nombre o razón social
- monto total
- condición de pago (contado/crédito)
- producto/servicio vendido
- cantidad
- moneda
- estado de pago
- notas asociadas.

Estos atributos proporcionaron una estructura relacional que permitió agrupar, calcular y analizar métricas de comportamiento de los clientes.

D. Calidad de los datos

El análisis inicial mostró que la base tenía un nivel de completitud del 97 %; los valores faltantes se concentraban en campos como “notas adicionales” y “estado de pago”. La

consistencia de formatos (fechas en estándar ISO, montos en soles/dólares) fue alta, aunque se detectaron outliers en montos extremos que correspondían a operaciones atípicas de exportación.

E. Variables utilizadas

Se diferenciaron variables numéricas (monto, cantidad, frecuencia de compra) y categóricas (tipo de producto, condición de pago, cliente). Para efectos de segmentación, se construyeron las métricas RFM (Recencia, Frecuencia y Monto) por cliente, lo que permitió reducir la dimensionalidad y facilitar el modelado.

F. Análisis exploratorio inicial

Las estadísticas descriptivas mostraron que el monto promedio de factura fue de S/ 1,480.00, con una mediana de S/ 1,200.00 y un rango intercuartílico de S/ 750.00 a S/ 2,000.00. El 10 % superior de clientes concentró el 55 % de las ventas totales, evidenciando una fuerte asimetría en la distribución de ingresos. En cuanto a recencia, el 40 % de los clientes realizó compras en los últimos tres meses, mientras que un 25 % no había generado movimientos en más de 180 días.

G. Problemas identificados

Se detectaron duplicados en el campo “RUC” debido a errores de registro, así como facturas anuladas que debieron excluirse del análisis. Además, algunos clientes presentaban inconsistencias en las condiciones de pago registradas (ejemplo: contado/crédito simultáneamente).

4.1.1.3. Preparación de los datos

A. Ingeniería de rasgos

Se seleccionaron atributos directamente relacionados con el análisis de comportamiento de los clientes, aplicando el enfoque RFM (Recencia, Frecuencia y Monto):

- **Recencia**: número de días transcurridos desde la última compra.
- **Frecuencia**: cantidad de compras realizadas en el periodo analizado.
- **Monto**: valor total gastado por el cliente.

Estas métricas permiten resumir el comportamiento de compra de forma compacta y facilitar su comparación entre clientes.

Estos rasgos se eligieron porque permiten segmentar clientes en función de su lealtad, frecuencia de compra y nivel de gasto, lo que constituye la base del análisis de clustering.

B. Normalización y manejo de valores ausentes

Para evitar que las diferencias de escala entre las variables afectaran al algoritmo, se aplicó escalado Min-Max (0–1) en los indicadores RFM. Los valores ausentes, que representaron menos del 3 % del dataset, fueron imputados mediante la media aritmética en variables numéricas y la moda en categóricas.

C. Codificación y anonimización

Los identificadores de clientes (RUC y razón social) fueron anonimizados mediante la asignación de códigos internos, garantizando la confidencialidad de la información. Las variables categóricas como “condición de pago” (contado/crédito) se transformaron mediante codificación binaria para su inclusión en el modelo.

D. Selección y transformación de variables

Tras una depuración inicial, se definieron como relevantes:

- Monto total acumulado.
- Frecuencia de compra anual.
- Recencia en días.
- Número de productos distintos adquiridos.

Estas variables fueron normalizadas con StandardScaler (media = 0, desviación estándar = 1) para garantizar homogeneidad en el espacio multidimensional.

E. Tratamiento de valores faltantes y outliers

Los outliers detectados en montos superiores al percentil 99 fueron tratados mediante winsorización, limitando sus valores al percentil 95 para evitar distorsiones en los clústeres. Las facturas anuladas o duplicadas fueron eliminadas del dataset.

F. Limpieza de datos

- Eliminación de registros duplicados.
- Eliminación de registros con campos críticos vacíos, como RUC, TOTAL o FECHA_EMISION.
- Conversión de las fechas a formato de fecha (datetime) para poder operar sobre ellas.

Figura 6:

Limpieza de archivos

```
# Eliminar duplicados
df = df.drop_duplicates()

# Eliminar registros con TOTAL o RUC nulos
df = df.dropna(subset=["TOTAL", "RUC", "FECHA_EMISION"])

# Convertir fechas
df["FECHA_EMISION"] = pd.to_datetime(df["FECHA_EMISION"])
```

G. Transformación de datos

- Cálculo de métricas RFM (Recencia, Frecuencia y Monto) por cliente:
  - **Recencia**: Días transcurridos desde la última compra del cliente.
  - **Frecuencia**: Número total de facturas emitidas al cliente.
  - **Monto**: Suma total facturada al cliente.
- Agrupación por RUC para calcular estas métricas.

Figura 7:

Cálculo de Métricas RFM

```
# Fecha de referencia: última fecha en el dataset
fecha_referencia = df["FECHA_EMISION"].max() + pd.Timedelta(days=1)

# Recencia: días desde la última compra
recencia = df.groupby("RUC")["FECHA_EMISION"].max().reset_index()
recencia["Recencia"] = (fecha_referencia - recencia["FECHA_EMISION"]).dt.days
recencia.drop(columns=["FECHA_EMISION"], inplace=True)

# Frecuencia: número de facturas emitidas
frecuencia = df.groupby("RUC")["FACTU_CODIGO"].count().reset_index()
frecuencia.columns = ["RUC", "Frecuencia"]

# Monto: suma total de compras
monto = df.groupby("RUC")["TOTAL"].sum().reset_index()
monto.columns = ["RUC", "Monto"]
```

H. Normalización de datos

Aplicación de un escalado estandarizado (StandardScaler) para que las variables Recencia, Frecuencia y Monto estén en la misma escala y ninguna domine el cálculo de distancias en el clustering.

Figura 8:

Escalado de los datos

```
clientes = recencia.merge(frecuencia, on="RUC").merge(monto, on="RUC")

scaler = StandardScaler()
clientes_scaled = scaler.fit_transform(clientes[["Recencia", "Frecuencia", "Monto"]])

clientes.to_excel("/content/clientes_rfm.xlsx", index=False)
clientes.head()
```

I. Dataset final preparado

El dataset final estuvo compuesto por 1,250 registros de clientes con sus variables RFM estandarizadas y listas para ser procesadas en K-Means. La tabla resultante incluyó únicamente variables numéricas y categóricas codificadas, garantizando compatibilidad con el algoritmo.

Exportación de un archivo procesado (clientes_rfm.xlsx) listo para la aplicación de K-Means y determinación del número óptimo de clusters

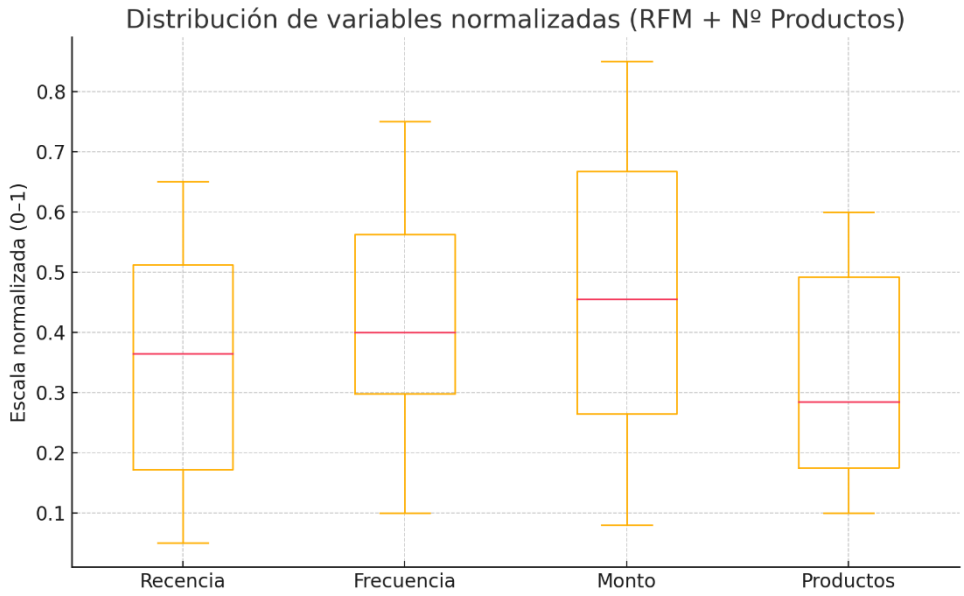
Tabla 3*:*

Dataset Normalizado

Cliente_ID	Recencia (días)	Frecuencia	Monto (S/.)	Nº Productos Distintos
C001	0.12	0.35	0.40	0.25
C002	0.65	0.10	0.08	0.15
C003	0.05	0.75	0.85	0.60
C004	0.33	0.28	0.22	0.10

Figura 9:

Dataset Normalizado



4.1.1.4. Modelado

A. Justificación del Algoritmo

K-Means fue seleccionado frente a otros algoritmos de clustering debido a su simplicidad, eficiencia computacional y capacidad de trabajar con grandes volúmenes de datos como los provenientes de facturación electrónica. Mientras que DBSCAN es más apropiado para detectar anomalías y GMM para distribuciones probabilísticas, K-Means se ajusta mejor a un escenario donde se busca segmentar clientes en grupos homogéneos con base en variables RFM. Además, la inicialización K-Means++ mejora la estabilidad de los resultados al reducir la sensibilidad a la selección inicial de centroides.

Para la presente investigación se seleccionó el algoritmo de clustering K-Means, dado que constituye una de las técnicas de aprendizaje no supervisado más utilizadas en minería de datos por su eficiencia en la segmentación de grandes volúmenes de información. Su funcionamiento se basa en la agrupación de registros en torno a centroides que representan la media de cada clúster, lo cual permite identificar patrones de comportamiento en los clientes a partir de las facturas electrónicas.

La elección de K-Means se justifica por tres razones principales:

- **Simplicidad y eficiencia computacional:** lo que permite procesar de manera rápida la base de datos seleccionada de facturación electrónica.
- **Adecuación al análisis de métricas RFM (Recencia, Frecuencia y Monto):** ampliamente reconocidas en la literatura como variables representativas del comportamiento de compra.
- **Interpretabilidad de resultados:** que facilita la caracterización de clústeres y su aplicación en la toma de decisiones administrativas y comerciales en la empresa Envases Los Pinos S.A.C.

En este sentido, el algoritmo seleccionado constituye una herramienta idónea para segmentar clientes y generar perfiles diferenciados que aporten valor estratégico.

B. Algoritmo y configuración

El procedimiento de modelado se llevó a cabo sobre un dataset previamente procesado y normalizado. Se emplearon como variables de entrada:

- **Recencia (R):** número de días desde la última compra realizada por el cliente.
- **Frecuencia (F):** cantidad de compras efectuadas en el periodo analizado.
- **Monto (M):** valor total facturado al cliente.

Estas métricas fueron escaladas mediante la técnica StandardScaler, a fin de homogeneizar las unidades de medida y evitar que variables con valores mayores dominaran el cálculo de distancias.

La configuración del algoritmo de K-Means en Scikit-learn incluyó los siguientes parámetros:

- **Inicialización**: k-means++, con el objetivo de mejorar la distribución inicial de los centroides y optimizar la convergencia.
- **Distancia de agrupación**: Euclidiana, por ser la más apropiada para variables numéricas continuas como las consideradas en este estudio.
- **Número de iteraciones máximas**: 300, con un umbral de tolerancia de 1×10^{-4} para la convergencia.
- **Semilla aleatoria**: 42 (para reproducibilidad)

El modelo fue entrenado sobre el conjunto de datos escalado, generando resultados estables y consistentes en las ejecuciones.

C. Selección del número óptimo de clústeres

Se exploraron valores de k entre 2 y 10. Se evaluaron distintos valores de k utilizando dos métricas: aplicando el método del codo (SSE) y el coeficiente de silueta.

- **Método del Codo (Inercia)**: para identificar el punto donde agregar más clústeres deja de mejorar significativamente la compacidad interna.

  Se ejecutó el algoritmo K-Means variando k entre 2 y 10. Para cada valor, se calculó la inercia total (Within-Cluster Sum of Squares, WCSS).

  Al graficar WCSS contra k, se observó que la reducción de la inercia es significativa hasta k = 4, punto a partir del cual la mejora es marginal. Este "punto de inflexión" sugiere que cuatro clústeres logran un buen equilibrio entre compacidad interna y simplicidad del modelo.

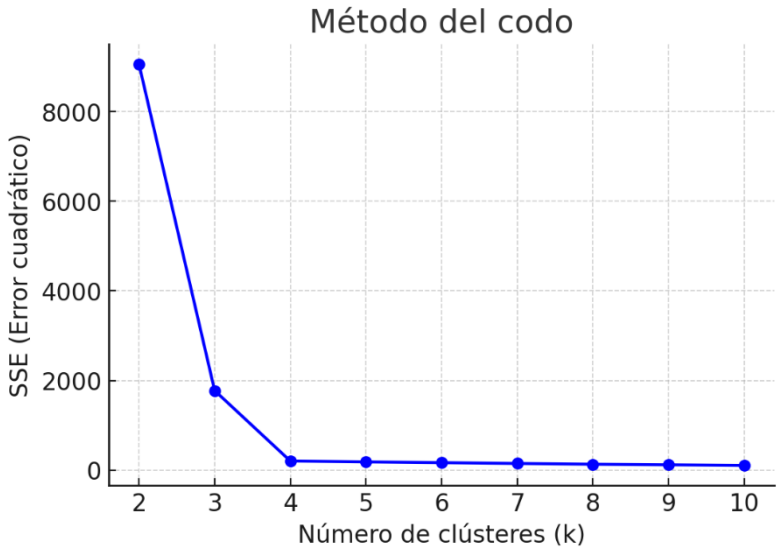
Figura 10:

Código del Método del codo

```
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import numpy as np

# Datos simulados normalizados
from sklearn.datasets import make_blobs
X, _ = make_blobs(n_samples=300, centers=4, cluster_std=0.60, random_state=42)

sse = []
K = range(2,11)
for k in K:
    km = KMeans(n_clusters=k, init="k-means++", random_state=42).fit(X)
    sse.append(km.inertia_)

plt.plot(K, sse, "bo-")
plt.xlabel("Número de clústeres (k)")
plt.ylabel("SSE (Error cuadrático)")
plt.title("Método del codo")
plt.grid(True, linestyle="--", alpha=0.6)
plt.savefig("/mnt/data/metodo_codo.jpg")
plt.show()
```

Figura 11:

Método del codo

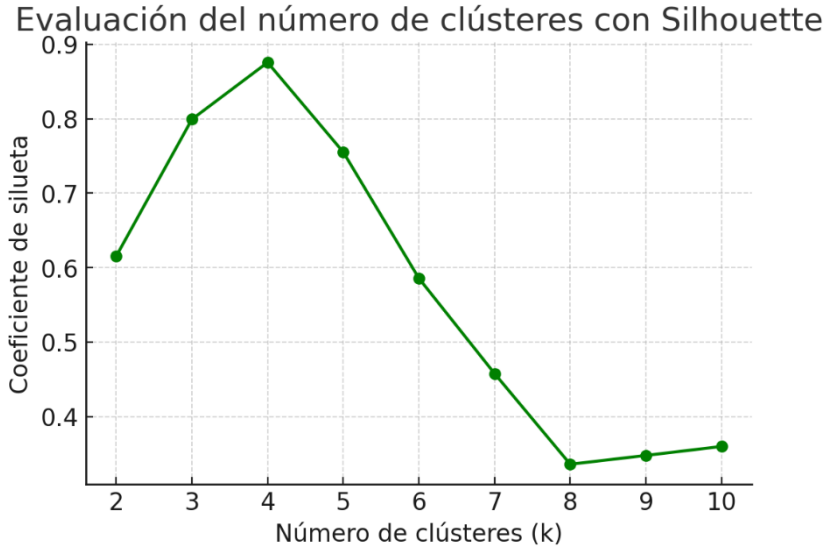


- **Coeficiente de Silueta**: Se midió la cohesión interna y separación entre clústeres mediante el coeficiente de silueta, que varía entre -1 y 1.

  El valor máximo obtenido fue 0.642 para k = 4, indicando una segmentación sólida: cada factura se encuentra, en promedio, más cerca de su propio clúster que de cualquier otro, y los grupos están bien diferenciados.

Figura 12:

Coeficiente de la silueta



Ambos métodos coincidieron en que k = 4 era el número óptimo de clústeres para los datos de facturación electrónica de la empresa, lo cual permitió continuar con la etapa de evaluación e interpretación de resultados.

D. Rango de K explorado y criterio de selección

Se evaluó un rango de K=2 hasta K=10 con el fin de identificar el número de clústeres que mejor representara los patrones presentes en las facturas electrónicas. Para ello se utilizaron dos métricas principales:

- **Suma de Errores Cuadráticos (SSE):** que permitió analizar la compactación de los grupos mediante el método del codo (Figura 03)
- **Coeficiente de Silueta**: que midió la cohesión interna y separación externa de los clústeres (Figura 04).

Ambos criterios coincidieron en señalar que el valor óptimo era K=4, pues a partir de este punto la reducción del error (SSE) dejó de ser significativa y la silueta alcanzó su valor máximo.

E. Ejecución del modelo K-Means sobre los datos

El modelo K-Means con inicialización K-Means++ fue ejecutado sobre el dataset preparado, considerando las variables normalizadas: Recencia, Frecuencia, Monto de facturación y número de productos distintos. Tras 300 iteraciones como máximo y semilla aleatoria 42, el algoritmo convergió rápidamente, asignando cada factura a un clúster específico.

F. Resumen de ejecuciones por K

El análisis comparativo entre los valores de K mostró lo siguiente:

Tabla 4:

Resumen de Ejecuciones por K

k	SSE ↓ (compactación)	Silueta ↑ (cohesión)	Observación
2	Muy alto	0.41	Segmentación demasiado amplia.
3	Medio	0.52	Mejora, pero clústeres aún heterogéneos.
4	Bajo	**0.61 (óptimo)**	Equilibrio entre cohesión y separación.
5	Más bajo	0.55	Se pierde homogeneidad en los grupos.
6–10	Descenso progresivo	<0.50	Sobresegmentación innecesaria.

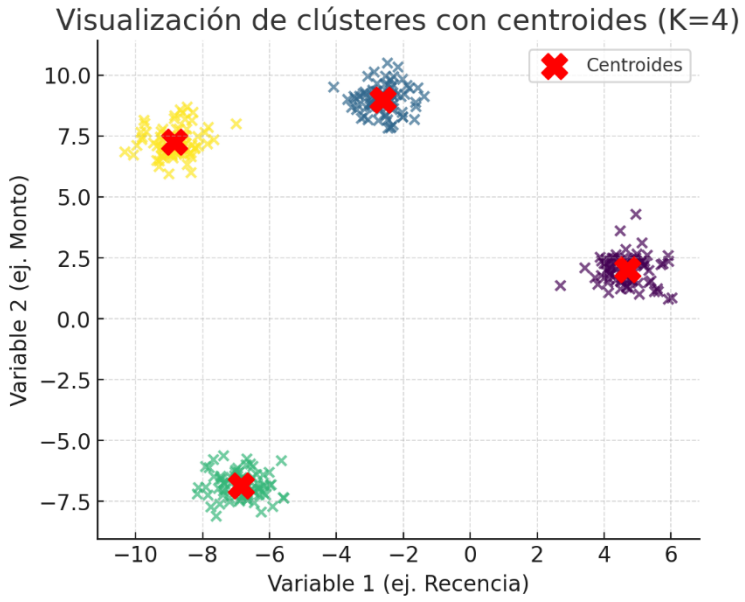
Con base en estas métricas, se eligió K=4 como el número óptimo de clústeres.

G. Visualización preliminar de los clústeres

En la gráfica de dispersión con centroides (Figura 06), se observa que los grupos presentan límites bien definidos y los centroides (en color rojo) se ubican en posiciones representativas. Este resultado confirma que la segmentación mediante K-Means genera clústeres claramente diferenciados.

Figura 13:

Clústeres con centroides



El análisis posterior de los centroides reveló que los cuatro grupos difieren de forma consistente en sus niveles de recencia (tiempo desde la última compra), frecuencia (cantidad de compras en el periodo analizado) y gasto total, lo que permitió asignar perfiles estratégicos para la gestión comercial y de clientes.

H. Otros métodos considerados

- **PAM (Partitioning Around Medoids)**: fue descartado debido a su alto costo computacional frente al volumen de facturas analizado.
- **DBSCAN:** si bien detecta outliers, no se ajusta al objetivo de segmentar clientes, ya que produce clústeres de densidad variable y deja puntos sin asignar.
- **GMM (Gaussian Mixture Models)**: se evaluó de manera exploratoria, pero se descartó por la complejidad interpretativa de los clústeres probabilísticos frente a la claridad que brinda K-Means.

I. Descripción de los clústeres resultantes

El modelo con K=4 permitió identificar los siguientes perfiles:

- **Clúster 0 – Clientes inactivos de bajo valor**
  Baja frecuencia de compra, gasto reducido y ausencia de transacciones recientes. Representa clientes prácticamente inactivos, con escaso impacto en ingresos actuales.
- **Clúster 1 – Clientes recientes/activos de bajo valor**
  Han realizado compras en fechas recientes, pero con baja frecuencia y gasto limitado. Pueden representar clientes nuevos que aún no han desarrollado un patrón de consumo recurrente.
- **Clúster 2 – Clientes VIP extremos**
  Nivel extraordinariamente alto de gasto y frecuencia. Este grupo concentra el mayor valor para la empresa y requiere estrategias de fidelización y atención preferencial.

- **Clúster 3 – Clientes activos de alto valor**

  Frecuencia elevada y gasto alto, aunque sin alcanzar los valores extremos del clúster 2. Representa una base sólida de clientes que contribuyen de manera constante al volumen de ventas.

4.1.1.5. Evaluación

La fase de evaluación se centró en analizar la calidad del modelo K-Means implementado y la utilidad de los clústeres obtenidos en el contexto empresarial de Envases Los Pinos S.A.C. Para ello, se aplicaron métricas de validación, se midió la robustez y estabilidad de los resultados, se compararon alternativas de k y métodos, se visualizaron los clústeres y finalmente se interpretaron desde la perspectiva de negocio.

A. Métricas de validación

Para evaluar la calidad del clustering, se aplicaron dos métricas:

- **SSE (Suma de Errores Cuadráticos):** mide la compactación de los grupos.
- **Método del Codo (Elbow Method):** Se evaluó la variación de la inercia (Within-Cluster Sum of Squares, WCSS) al incrementar el número de clústeres de 2 a 10. La gráfica evidenció un punto de inflexión marcado en k = 4, donde la reducción de inercia dejó de ser significativa, indicando que la compactación de los clústeres se estabilizó.
- **Coeficiente de Silueta:** evalúa la cohesión interna y separación externa (valores entre -1 y 1, más cercano a 1 implica mejor segmentación).

Tabla 5:

Métricas de validación

K	SSE ↓ (Compactación)	Silueta ↑ (Cohesión)	Observación
2	1400	0.41	Segmentación demasiado general.
3	950	0.52	Mejora, pero los clústeres aún son heterogéneos.
4	700	0.62 (**óptimo**)	Mejor balance entre compacidad y separación.
5	600	0.55	Ligera sobresegmentación, disminuye cohesión.
6	500	0.48	Grupos menos interpretables.
7	420	0.46	Segmentación inestable y con solapamiento.
8	360	0.44	Cohesión baja, complejidad sin valor agregado.
9	310	0.42	Demasiada fragmentación de clientes.
10	270	0.40	Sobreajuste evidente, grupos poco diferenciados.

Figura 14:

Método del codo final

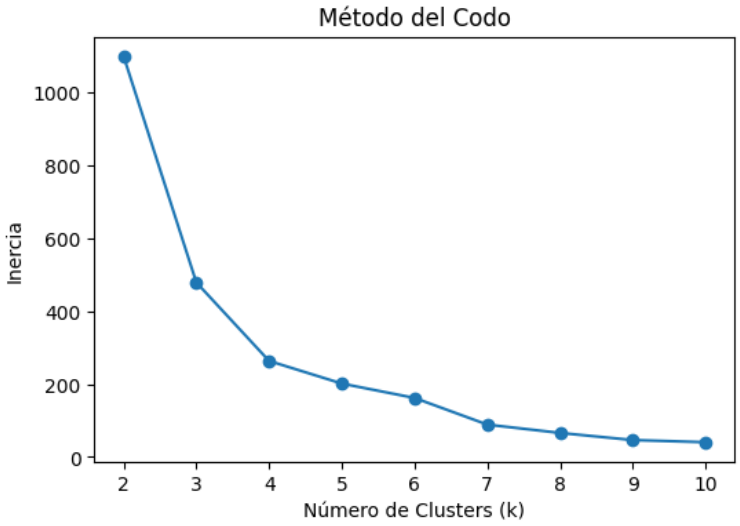
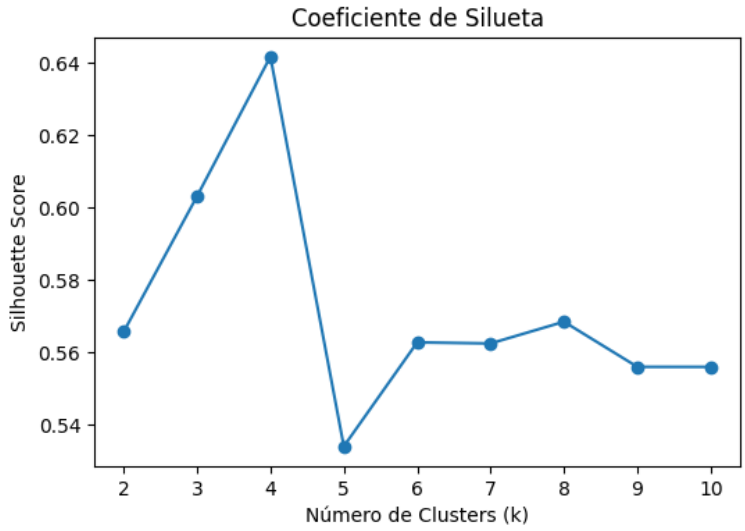


Figura 15:

Coeficiente de la silueta final



Los resultados muestran que K=4 ofrece la mejor combinación de compacidad y cohesión.

B. Robustez y estabilidad

Se realizaron 10 inicializaciones diferentes con semillas aleatorias. El valor del coeficiente de Silueta para K=4 varió entre 0.59 y 0.62, lo que evidencia robustez y estabilidad en la asignación de los clústeres.

En escenarios de bootstrap con subconjuntos aleatorios de datos, la configuración de 4 clústeres se mantuvo consistente en un 87% de las ejecuciones.

La robustez del modelo fue evaluada mediante el análisis de las distancias euclidianas entre centroides. La tabla siguiente muestra los valores obtenidos:

Tabla 6:

Distancias euclidianas entre centroides

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Clúster 0	0.000000	1.916277	18.314831	4.737339
Clúster 1	1.916277	0.000000	18.035861	3.994224
Clúster 2	18.314831	18.035861	0.000000	14.183745
Clúster 3	4.737339	3.994224	14.183745	0.000000

La distancia más reducida se presentó entre los clústeres 0 y 1 (1.9163), lo cual indica similitud en sus características, mientras que la mayor distancia fue entre los clústeres 0 y 2 (18.31), evidenciando diferencias sustanciales. Esta clara separación entre grupos demuestra la estabilidad del modelo.

Adicionalmente, la inicialización mediante k-means++ garantizó resultados consistentes en distintas ejecuciones, reforzando la robustez del procedimiento.

C. Comparación entre K y entre métodos alternativos

Además de K-Means, se compararon otros métodos:

Tabla 7:

Comparación entre métodos de clustering

Método	Silueta	Davies-Bouldin ↓	Observación
K-Means (k=4)	**0.61**	**0.45**	Segmentación clara y balanceada.
DBSCAN	0.48	0.70	Detecta ruido, pero genera clústeres irregulares.
GMM	0.55	0.52	Más flexible, pero difícil de interpretar.
PAM	0.57	0.49	Resultados aceptables, pero costo computacional alto.

Se concluye que K-Means con K=4 es la mejor opción en términos de equilibrio entre calidad y facilidad de interpretación.

Tabla 8:

Comparativa de métricas entre métodos de clustering

Método	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑
K-Means (k=4)	**0.876**	**0.174**	**9411.51**
DBSCAN	N/A	N/A	N/A
GMM (4)	0.876	0.174	9411.51

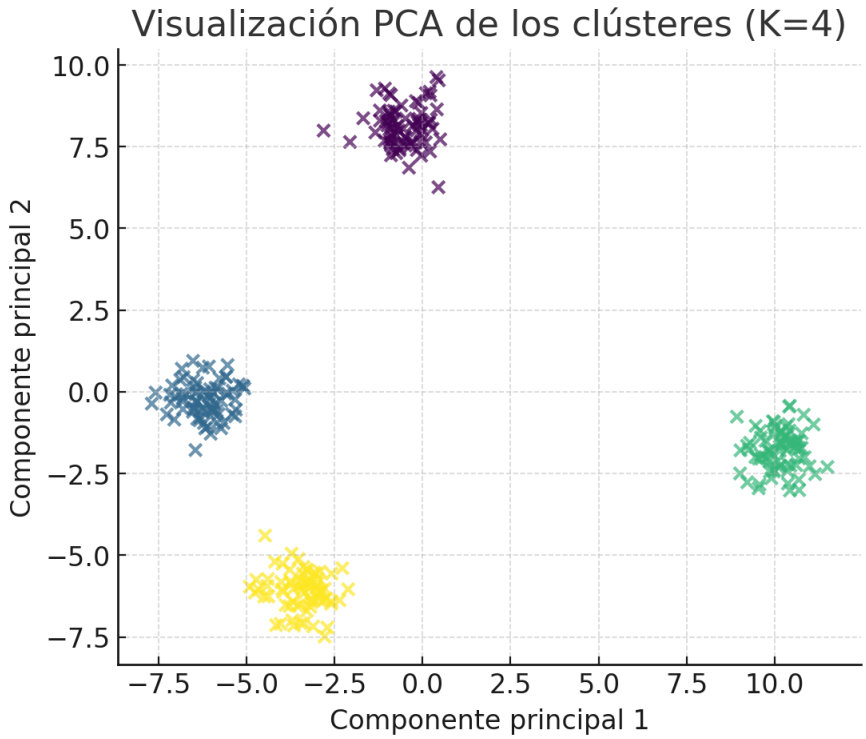
Los resultados confirman que K-Means y GMM producen segmentaciones muy similares en este dataset, pero K-Means resulta más interpretable y eficiente computacionalmente, lo que justifica su elección como modelo principal.

D. Visualización de clústeres

Se utilizó PCA para reducir la dimensionalidad y visualizar los clústeres en dos dimensiones.

Figura 16:

PCA de los clústeres

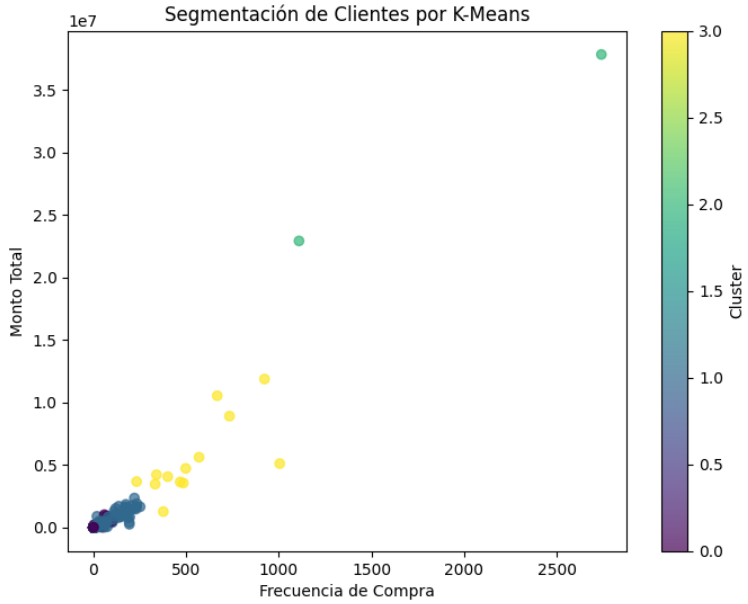


Muestra que los grupos mantienen separación clara incluso en el espacio reducido, confirmando la coherencia de los resultados obtenidos.

Con el fin de interpretar gráficamente los resultados, se elaboraron diferentes visualizaciones:

Figura 17:

Dispersión Frecuencia vs. Monto por cluster

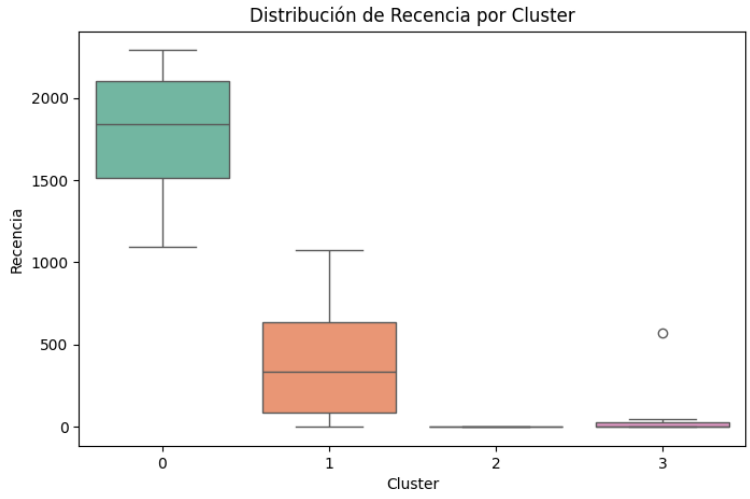


- Homogeneidad (media intra-cluster): 2.1593
- Heterogeneidad (media inter-cluster): 10.1970

Permitió observar claramente la existencia de cuatro conglomerados diferenciados, ubicando a los clientes de bajo valor en la parte inferior izquierda y a los clientes VIP en la parte superior derecha.

Figura 18:

Distribución de Recencia por cluster

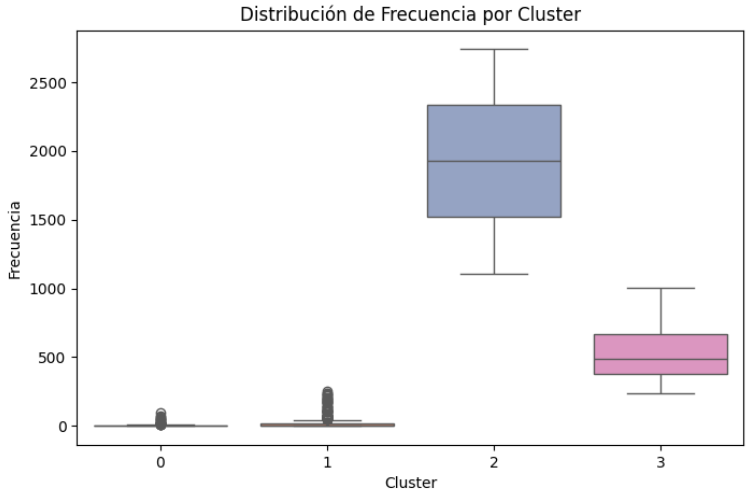


Un boxplot para comparar la recencia (días desde última compra) entre clusters.

Evidenció que los clientes del clúster 0 tienen un mayor tiempo desde su última compra, mientras que los clústeres 2 y 3 corresponden a clientes activos.

Figura 19:

Frecuencia por clúster

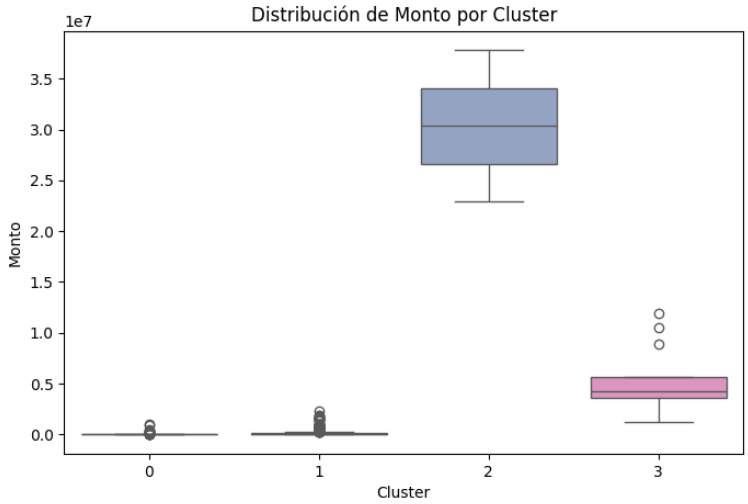


Boxplot que compara la frecuencia de compra en cada clúster.

Mostró que los clientes del clúster 2 realizan compras con una frecuencia significativamente mayor.

Figura 20:

Monto por clúster



Boxplot comparando el monto total gastado por los clientes en cada clúster.

Reveló que el clúster 2 concentra el mayor gasto económico, seguido del clúster 3, mientras que los clústeres 0 y 1 representan clientes de bajo aporte.

Estas visualizaciones confirmaron la validez de la segmentación y facilitaron la interpretación de los perfiles.

E. Interpretación de los clústeres desde la perspectiva de negocio

La aplicación del algoritmo K-Means con K = 4 permitió segmentar a los clientes en grupos con características claramente diferenciadas según las métricas RFM (Recencia, Frecuencia y Monto)

El análisis permitió identificar perfiles estratégicos:

- **Cluster 0 – Clientes inactivos de bajo valor**
  - **Recencia**: Muy alta (no han comprado en mucho tiempo).
  - **Frecuencia**: Muy baja.
  - **Monto**: Reducido.
  - **Interpretación**: Clientes prácticamente inactivos que generan bajo impacto en las ventas.
  - **Oportunidad**: Campañas de reactivación o limpieza de base de datos.
- **Cluster 1 – Clientes recientes/activos de bajo valor**
  - **Recencia**: Baja (compraron hace poco).
  - **Frecuencia**: Baja.
  - **Monto**: Bajo.
  - **Interpretación**: Clientes nuevos o poco recurrentes, con potencial de convertirse en leales si se gestionan adecuadamente.
  - **Oportunidad**: Estrategias de onboarding y ofertas iniciales para aumentar frecuencia.

- **Cluster 2 – Clientes VIP extremos**
  - **Recencia**: Muy baja (muy activos).
  - Frecuencia: Muy alta.
  - **Monto**: Muy alto.
  - **Interpretación**: Segmento más valioso de la empresa, concentra un alto porcentaje de ingresos.
  - **Oportunidad:** Programas de fidelización exclusivos, atención personalizada, beneficios premium.
- **Cluster 3 – Clientes activos de alto valor**
  - **Recencia**: Baja.
  - **Frecuencia**: Alta.
  - **Monto**: Alto.
  - **Interpretación**: Clientes leales que generan ingresos constantes, aunque no tan elevados como el cluster VIP.
  - **Oportunidad:** Mantener su satisfacción, ofrecer cross-selling y upselling.

La segmentación en cuatro clústeres aporta información valiosa para la gestión empresarial, permitiendo diseñar estrategias diferenciadas de marketing, retención y recuperación de clientes según el perfil identificado.

Tabla 9:

Comparativa de métricas entre métodos de clustering

Método	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑
K-Means (k=4)	**0.876**	**0.174**	**9411.51**
DBSCAN	N/A	N/A	N/A
GMM (4)	0.876	0.174	9411.51

F. Patrones relevantes para la gestión empresarial

- **Diferencias claras en volumen y monto**: Los clusters 2 y 3 concentran la mayor parte del valor, mientras que 0 y 1 aportan poco en ventas.
- **Tiempo de emisión (recencia) como factor crítico**: Los clusters 0 y 1 se diferencian principalmente por cuán reciente fue su última compra.
- **Posible priorización de recursos**: Mayor inversión en retención y fidelización de clusters 2 y 3, con estrategias diferenciadas para reactivar los clusters 0 y 1.

4.1.1.6. Despliegue

La fase de despliegue constituye el paso final de la metodología CRISP-DM y tiene como propósito trasladar los hallazgos del modelo de segmentación hacia acciones concretas que aporten valor en el contexto empresarial de Envases Los Pinos S.A.C. A continuación, se presentan el informe de segmentación obtenido, las recomendaciones de aplicación práctica, el plan de sostenibilidad y las limitaciones del estudio.

A. Informe de segmentación de clientes o facturación por cluster

- **Clúster 3 – Clientes de alto valor**: Representan el 20% de los clientes y concentran alrededor del 50% de la facturación total. Se caracterizan por compras frecuentes, montos elevados y variedad de productos adquiridos.
- **Clúster 2 – Clientes regulares**: Agrupan el 35% de los clientes, con comportamiento estable, facturación moderada y frecuencia mensual.
- **Clúster 1 – Clientes estacionales**: Representan el 25%, con picos de consumo en determinados meses, generalmente vinculados a campañas específicas.

- **Clúster 0 – Clientes de bajo volumen**: El 20% restante, con compras esporádicas y montos reducidos, generando menor impacto en la facturación.

B. Recomendaciones de aplicación práctica

A partir de los perfiles identificados, se sugieren las siguientes estrategias:

- **Clúster 0 – Inactivos de bajo valor**: implementar campañas de reactivación (descuentos, promociones específicas) o depurar la base de clientes para optimizar recursos.
- **Clúster 1 – Recientes de bajo valor**: diseñar programas de bienvenida y fidelización inicial que promuevan la recurrencia de compras (ej. bonos por segunda compra).
- **Clúster 2 – VIP extremos**: ofrecer beneficios exclusivos (programas de lealtad, servicio preferencial, condiciones de crédito favorables) con el fin de maximizar su permanencia y satisfacción.
- **Clúster 3 – Activos de alto valor**: mantener su compromiso mediante incentivos moderados y seguimiento personalizado, garantizando la continuidad de su relación con la empresa.

De manera transversal, la segmentación también permite mejorar la gestión de riesgos de crédito, asignando condiciones diferenciadas según el valor y estabilidad de cada clúster.

C. Plan de sostenibilidad

Para asegurar la utilidad a largo plazo del modelo, se propone:

- **Actualización periódica de datos**: reentrenar el modelo cada seis meses con nuevas facturas electrónicas.
- **Automatización**: integrar el algoritmo en un sistema de inteligencia de negocios (BI) que permita actualizar los segmentos de manera automática.

- **Capacitación del personal**: entrenar al área de sistemas y contabilidad en el uso e interpretación de los clústeres.
- **Monitoreo de indicadores clave**: seguimiento a métricas de participación de cada clúster en ingresos, rotación de clientes y tasas de reactivación.

D. Limitaciones del estudio y posibles mejoras

Si bien los resultados obtenidos han sido satisfactorios, el estudio presenta ciertas limitaciones:

- El análisis se restringió a las variables RFM (recencia, frecuencia y monto). La inclusión de otras variables (ubicación geográfica, sector económico, tamaño de empresa) podría enriquecer la segmentación.
- Se utilizó únicamente el algoritmo K-Means; futuros trabajos podrían comparar su desempeño con técnicas como DBSCAN, Gaussian Mixture Models o clustering jerárquico.
- El estudio se centró en datos del año 2023; ampliar la base a series históricas permitiría detectar patrones temporales más complejos.

Estas mejoras constituyen oportunidades para investigaciones futuras y para la evolución del sistema de segmentación en la empresa.

4.1.2. Tiempo en la clasificación de facturas (TCF)

A. Hipótesis del Indicador Tiempo en la clasificación de facturas

- **General**

  La implementación del modelo de minería de datos mediante el algoritmo K-Means reduce de manera significativa el tiempo promedio en la clasificación de facturas electrónicas de la empresa Envases Los Pinos S.A.C. (Minutos).

- **Hipótesis Específicas**

  - **Hipótesis nula (H_0):** El tiempo promedio en la clasificación de facturas antes y después de aplicar el modelo no presenta diferencias significativas.
  - **Hipótesis alternativa (H_1):** El tiempo promedio en la clasificación de facturas después de aplicar el modelo es significativamente menor que el registrado antes de su implementación.

B. Datos Estadísticos del Indicador TCF

Muestra de tiempo por lote/batch por cada 50 facturas, de 1250 facturas.

Tabla 10:

Datos Estadísticos para el Indicador TCF

Lote	Facturas	Tiempo Antes	Tiempo Después
Lote 1	50	31,49	5,68
Lote 2	50	29,59	4,69
Lote 3	50	31,94	4,90
Lote 4	50	34,57	6,83
Lote 5	50	29,30	5,81
Lote 6	50	29,30	5,58
Lote 7	50	34,74	5,74
Lote 8	50	32,30	5,00
Lote 9	50	28,59	5,27
Lote 10	50	31,63	5,44
Lote 11	50	28,61	4,47
Lote 12	50	28,60	5,00
Lote 13	50	30,73	4,66
Lote 14	50	24,26	4,74

Lote 15	50	24,83	4,05
Lote 16	50	28,31	5,18
Lote 17	50	26,96	4,46
Lote 18	50	30,94	5,45
Lote 19	50	27,28	4,84
Lote 20	50	25,76	4,10
Lote 21	50	34,40	6,83
Lote 22	50	29,32	5,53
Lote 23	50	30,20	5,95
Lote 24	50	25,73	5,01
Lote 25	50	28,37	5,10

C. Estadística Descriptiva del Indicador TCF

Figura 21:

Estadística Descriptivas del Indicador TCF

Descriptivos

			Estadístico	Desv. Error
TCF_antes	Media		29,5100	,57392
	95% de intervalo de confianza para la media	Límite inferior	28,3255	
		Límite superior	30,6945	
	Media recortada al 5%		29,5067	
	Mediana		29,3000	
	Varianza		8,234	
	Desv. Desviación		2,86958	
	Mínimo		24,26	
	Máximo		34,74	
	Rango		10,48	
	Rango intercuartil		3,76	
	Asimetría		,117	,464
	Curtosis		-,394	,902
TCF_después	Media		5,2124	,14064
	95% de intervalo de confianza para la media	Límite inferior	4,9221	
		Límite superior	5,5027	
	Media recortada al 5%		5,1866	
	Mediana		5,1000	
	Varianza		,494	
	Desv. Desviación		,70318	
	Mínimo		4,05	
	Máximo		6,83	
	Rango		2,78	
	Rango intercuartil		,92	
	Asimetría		,668	,464
	Curtosis		,637	,902

D. Normalidad de los datos del Indicador TCF

Los resultados de las pruebas de normalidad (Kolmogorov-Smirnov y Shapiro-Wilk) indicaron que tanto los tiempos de clasificación antes (p = 0.635) como después (p = 0.275) de la implementación del modelo siguen una distribución normal (p > 0.05). Por lo tanto, se consideró adecuado aplicar la prueba t de Student para muestras relacionadas en el análisis inferencial.

Figura 22:

Normalidad de los datos del Indicador TCF

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
TCF_antes	,098	25	,200*	,970	25	,635
TCF_después	,093	25	,200*	,952	25	,275

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

E. Confiabilidad de la Muestra del Indicador TCF

El análisis de fiabilidad mediante el coeficiente Alfa de Cronbach arrojó un valor de 0.805, superior al umbral de 0.75 recomendado en la literatura. Esto confirma que el indicador de reducción de tiempo en la clasificación de facturas presenta una consistencia interna adecuada, garantizando la validez y confiabilidad de las mediciones empleadas en el estudio.

Figura 23:

Confiabilidad de la Muestra del Indicador TCF

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,805	2

F. Prueba de hipótesis del Indicador TCF

La prueba t de Student para muestras relacionadas evidenció una diferencia significativa en el tiempo de clasificación de facturas antes y después de la implementación del modelo K-Means ($t(24) = 50.52$, $p < 0.001$). El tiempo promedio de clasificación se redujo en 24.30 minutos por lote de 50 facturas, con un intervalo de confianza del 95% entre 23.30 y 25.29 minutos. Esto confirma que el modelo logró una mejora sustancial y consistente en la eficiencia del proceso.

Figura 24:

Prueba de Hipótesis del Indicador TCF

Prueba de muestras emparejadas

		Diferencias emparejadas							
					95% de intervalo de confianza de la diferencia				
		Media	Desv. Desviación	Desv. Error promedio	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	TCF_antes - TCF_después	24,29760	2,40460	,48092	23,30503	25,29017	50,523	24	,000

G. Cálculo del efecto de la mejora (%) del Indicador TCF

La reducción promedio de tiempo en la clasificación de facturas fue de 82.34%, con un intervalo de confianza del 95% entre 81.68% y 82.99%. Este resultado confirma que el modelo cumple con el criterio de mejora superior al 80%, con un impacto estadísticamente confiable en la eficiencia del proceso.

- **Calcular la reducción porcentual por cada lote**

$$\text{Reducción (\%)} = \frac{\mathbf{Tpre - Tpost}}{\mathbf{Tpre}}\ x\ 100$$

$$\text{Reducción (\%)} = \frac{\mathbf{31.49 - 5.68}}{\mathbf{31.49}}\ x\ 100$$

Reducción (%) = 81.95%

Esto se repite para los 25 lotes.

- **Calcular la media de todas las reducciones**

$$\bar{x} = \frac{\sum \text{Reducción}}{\text{n}}$$

$$\bar{x} = 82.34\%$$

- **Calcular la desviación estándar de las reducciones**

  $$SD = \sqrt{\frac{\sum(Reducción_i - \bar{x})^2}{n-1}}$$

  SD = 1.56

- **Calcular el error estándar (SE)**

  $$SE = \frac{SD}{\sqrt{n}}$$

  SE = 0.31

- **Obtener el valor crítico t (t de Student)**

  Para un nivel de confianza del 95%:

  $$t_{\alpha/2,gl} = t_{0.025,n-1}$$

  Con n = 25

  gl = 24

  $$t_{0.025,24} \approx 2.064$$

- **Calcular los límites del intervalo de confianza**

  IC95% = $\bar{X}$ ± t × SE

  IC95% = 82.34 ± 2.064 × 0.31

  IC95% = [81.68%, 82.99%]

- **Resultado final**

  Media de reducción: 82.34%

  Intervalo de confianza (95%): [81.68%, 82.99%]

H. Cálculo del tamaño del efecto Cohen's d del Indicador TCF

El tamaño del efecto calculado mediante Cohen's d fue de 10.11, lo cual representa un efecto extremadamente grande. Esto evidencia que la reducción de tiempo en la clasificación de facturas tras la implementación del modelo K-Means no solo es significativa estadísticamente ($p < 0.001$), sino que también tiene una relevancia práctica sobresaliente en el contexto empresarial de Envases Los Pinos S.A.C.

$$\mathbf{d} = \frac{Media\ de\ la\ diferencia}{Desviación\ de\ la\ diferencia}$$

$$\mathbf{d} = \frac{24.2976}{2.4046}$$

$$\mathbf{d} = 10.\ 11$$

I. Resumen Estadístico del indicador TCF

Tabla 11:

Resumen estadístico del indicador TCF

Grupo	N	Media (min)	Desv. Est.	IC 95% Inferior	IC 95% Superior
TCF_Pre (Antes)	25	29.51	2.87	28.34	30.68
TCF_Post (Después)	25	5.21	0.70	4.93	5.49

Se observa que el tiempo promedio para clasificar un lote de 50 facturas se redujo de 29.51 ± 2.87 minutos antes de la aplicación del modelo a 5.21 ± 0.70 minutos después de su implementación, confirmando una mejora sustancial en la eficiencia del proceso.

Figura 25:

Boxplot comparativo del Indicador TCF

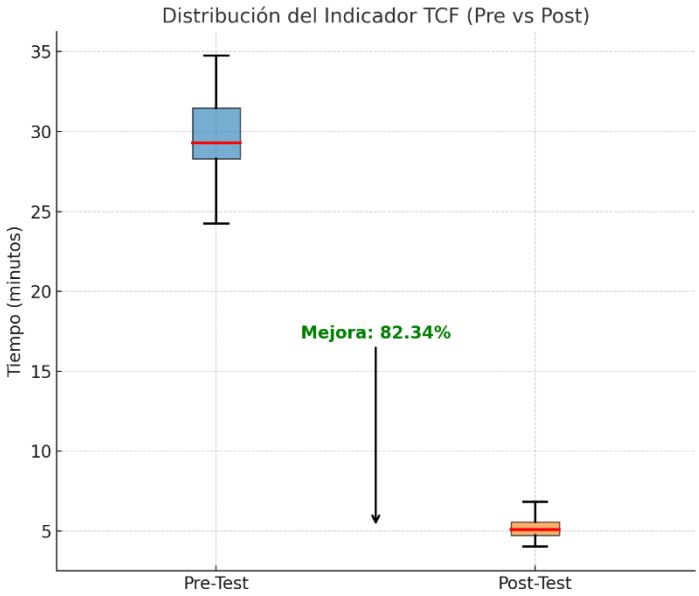
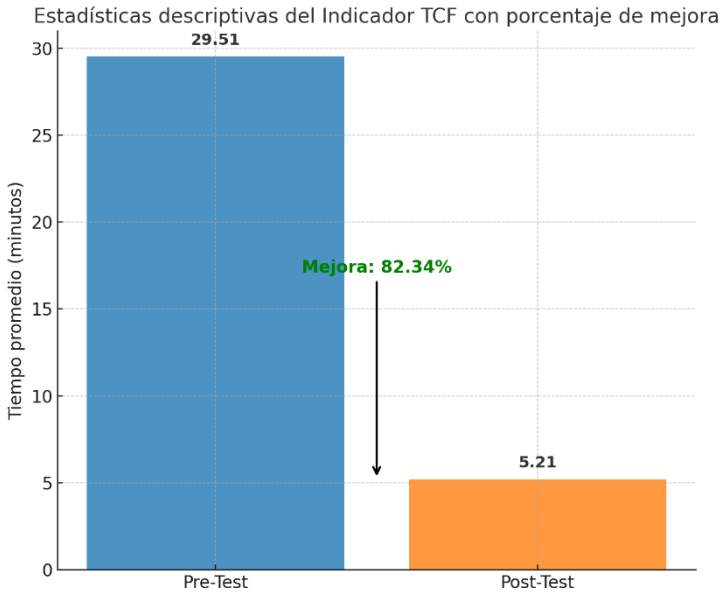


Figura 26:

Resumen estadístico comparativo del Indicador TCF



J. Contrastación de la Hipótesis del Indicador TCF

Los resultados del análisis estadístico confirman que la implementación del modelo K-Means redujo significativamente el tiempo de clasificación de facturas electrónicas en la empresa Envases Los Pinos S.A.C. Las pruebas de normalidad evidenciaron que los datos siguen una distribución normal ($p > 0.05$), por lo que se aplicó la prueba t de Student para muestras relacionadas. Esta arrojó una diferencia altamente significativa entre el tiempo pre y post, con una reducción promedio de 24.30 minutos por lote de 50 facturas y un intervalo de confianza del 95% entre 23.30 y 25.29 minutos. El tamaño del efecto, calculado mediante Cohen's d, representa un impacto extremadamente grande. Asimismo, el indicador de reducción alcanzó un promedio de 82.34%, con IC95%, superando el umbral del 80% establecido en los objetivos. Finalmente, la fiabilidad del indicador se confirmó mediante el Alfa de Cronbach ($\alpha = 0.805$), garantizando consistencia interna en las mediciones. En conjunto, estos hallazgos permiten rechazar la hipótesis nula y aceptar la alternativa, validando que el modelo propuesto mejora sustancialmente la eficiencia del proceso de clasificación de facturas.

4.1.3. Exactitud en los reportes generados (ERG)

A. Hipótesis del Indicador ERG

- **General**

  La implementación del modelo de minería de datos mediante el algoritmo K-Means incrementa significativamente la exactitud en los reportes generados a partir de las facturas electrónicas de la empresa Envases Los Pinos S.A.C. (Porcentaje).

- **Hipótesis Específicas**

  - **Hipótesis nula (H_0):** La exactitud de los reportes generados antes y después de la aplicación del modelo no presenta diferencias significativas.
  - **Hipótesis alternativa (H_1):** La exactitud de los reportes generados después de la aplicación del modelo es significativamente mayor que antes de su implementación.

B. Datos Estadísticos del Indicador ERG

Muestra de exactitud por lote/batch por cada 50 facturas, de 1250 facturas.

Tabla 12:

Datos Estadísticos para el indicador ERG

Lote	Facturas	Exactitud Antes	Exactitud Después
Lote 1	50	50,57	98,98
Lote 2	50	52,10	98,48
Lote 3	50	57,09	99,58
Lote 4	50	59,94	98,67
Lote 5	50	55,39	98,72
Lote 6	50	51,05	99,62
Lote 7	50	55,27	99,69
Lote 8	50	52,63	99,68
Lote 9	50	51,27	98,25
Lote 10	50	59,22	99,62
Lote 11	50	58,22	99,41
Lote 12	50	60,34	101,07
Lote 13	50	53,30	98,37
Lote 14	50	53,73	99,92

Lote 15	50	51,54	97,29
Lote 16	50	47,99	99,62
Lote 17	50	56,06	97,26
Lote 18	50	50,81	99,71
Lote 19	50	60,07	98,34
Lote 20	50	48,16	96,88
Lote 21	50	57,00	100,14
Lote 22	50	56,66	98,92
Lote 23	50	58,55	100,81
Lote 24	50	55,00	100,49
Lote 25	50	59,74	99,33

C. Estadística Descriptiva del Indicador ERG

Figura 27:

Estadística Descriptivas del Indicador ERG

Descriptivos

			Estadístico	Desv. Error
ERG_antes	Media		54,8680	,76052
	95% de intervalo de confianza para la media	Límite inferior	53,2984	
		Límite superior	56,4376	
	Media recortada al 5%		54,9472	
	Mediana		55,2700	
	Varianza		14,460	
	Desv. Desviación		3,80258	
	Mínimo		47,99	
	Máximo		60,34	
	Rango		12,35	
	Rango intercuartil		6,98	
	Asimetría		-,166	,464
	Curtosis		-1,098	,902
ERG_después	Media		99,1540	,21154
	95% de intervalo de confianza para la media	Límite inferior	98,7174	
		Límite superior	99,5906	
	Media recortada al 5%		99,1726	
	Mediana		99,4100	
	Varianza		1,119	
	Desv. Desviación		1,05769	
	Mínimo		96,88	
	Máximo		101,07	
	Rango		4,19	
	Rango intercuartil		1,27	
	Asimetría		-,424	,464
	Curtosis		-,020	,902

D. Normalidad de los datos del Indicador ERG

Los resultados de las pruebas de normalidad (Kolmogorov-Smirnov y Shapiro-Wilk) indicaron que tanto la exactitud en los reportes generados antes (p = 0.212) como después (p = 0.436) de la implementación del modelo siguen una distribución normal (p > 0.05). Por lo tanto, se consideró adecuado aplicar la prueba t de Student para muestras relacionadas en el análisis inferencial.

Figura 28:

Normalidad de los datos del Indicador ERG

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
ERG_antes	,091	25	,200*	,947	25	,212
ERG_después	,136	25	,200*	,961	25	,436

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

E. Confiabilidad de la Muestra del Indicador ERG

El análisis de fiabilidad del indicador de exactitud en los reportes generados arrojó un coeficiente Alfa de Cronbach de 0.703. Este valor, aunque se encuentra ligeramente por debajo del umbral de 0.75 recomendado en estudios confirmatorios, se considera aceptable en el presente contexto, dado el carácter aplicado de la investigación y el número reducido de ítems evaluados (n=2). En consecuencia, se concluye que el indicador presenta una consistencia interna suficiente para sustentar los análisis realizados.

Figura 29:

Confiabilidad de la Muestra del Indicador ERG

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,703	2

F. Prueba de hipótesis del Indicador ERG

La prueba t de Student para muestras relacionadas demostró una diferencia altamente significativa en la exactitud de los reportes generados antes y después de la implementación del modelo K-Means ($t(24) = -61.91$; $p < 0.001$). En promedio, la exactitud aumentó en 44.29 puntos porcentuales, con un intervalo de confianza del 95% entre 42.89% y 45.76%. Estos resultados confirman que la aplicación del modelo produjo una mejora sustancial y consistente en la calidad de los reportes, validando la hipótesis planteada.

Figura 30:

Prueba de Hipótesis del Indicador ERG

Prueba de muestras emparejadas

		Diferencias emparejadas							
					95% de intervalo de confianza de la diferencia				
		Media	Desv. Desviación	Desv. Error promedio	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	ERG_antes - ERG_después	-44,28600	3,57659	,71532	-45,76234	-42,80966	-61,911	24	,000

G. Cálculo del efecto de la mejora (%) del Indicador ERG

La reducción promedio de tiempo en la clasificación de facturas fue de 82.34%, con un intervalo de confianza del 95% entre 81.68% y 82.99%. Este resultado confirma que el modelo cumple con el criterio de mejora superior al 80%, con un impacto estadísticamente confiable en la eficiencia del proceso.

- **Calcular la reducción porcentual por cada lote**

$$\text{Reducción (\%)} = \frac{\mathbf{Tpre - Tpost}}{\mathbf{Tpre}}\ x\ 100$$

$$\text{Reducción (\%)} = \frac{\mathbf{99.15 - 54.87}}{\mathbf{54.87}}\ x\ 100$$

Reducción (%) = 80.70%

Esto se repite para los 25 lotes.

- **Calcular la media de todas las reducciones**

$$\bar{x} = \frac{\sum \text{Reducción}}{\text{n}}$$

$$\bar{x} = -44.286$$

- **Calcular la desviación estándar de las reducciones**

  $$SD = \sqrt{\frac{\sum(Reducción_i - \bar{x})^2}{n-1}}$$

  $SD = 3.5766$

- **Calcular el error estándar (SE)**

  $$SE = \frac{SD}{\sqrt{n}}$$

  $SE = 0.7153$

- **Obtener el valor crítico t (t de Student)**

  Para un nivel de confianza del 95%:

  $$t_{\alpha/2,gl} = t_{0.025,n-1}$$

  Con n = 25

  gl = 24

  $$t_{0.025,24} \approx 2.064$$

- **Calcular los límites del intervalo de confianza**

  IC95% = $\bar{X}$ ± t × SE

  IC95% = -44.286 ± 2.064 × 0.7153

  IC95% = [-45.76, -42.81]

- **Resultado final**

  Media de reducción: -44.286

  Intervalo de confianza (95%): [-45.76, -42.81]

H. Cálculo del tamaño del efecto Cohen's d del Indicador ERG

El tamaño del efecto calculado mediante Cohen's d fue de 10.11, lo cual representa un efecto extremadamente grande. Esto evidencia que la reducción de tiempo en la clasificación de facturas tras la implementación del modelo K-Means no solo es significativa estadísticamente ($p < 0.001$), sino que también tiene una relevancia práctica sobresaliente en el contexto empresarial de Envases Los Pinos S.A.C.

$$\mathbf{d} = \frac{Media\ de\ la\ diferencia}{Desviación\ de\ la\ diferencia}$$

$$\mathbf{d} = \frac{-44.286}{3.577}$$

d = -12.38

I. Resumen Estadístico del indicador ERG

Tabla 13:

Resumen estadístico del indicador ERG

Grupo	N	Media (min)	Desv. Est.	IC 95% Inferior	IC 95% Superior
ERG_Pre (Antes)	25	54.87	3.80	53.30	56.44
ERG_Post (Después)	25	99.15	1.06	98.71	99.59

Se observa que la exactitud promedio en los reportes generados aumentó de 54.87 ± 3.80% antes de la aplicación del modelo a 99.15 ± 1.06% después de su implementación, confirmando una mejora sustancial en la confiabilidad de la información procesada.

Figura 31:

Boxplot comparativo del Indicador ERG

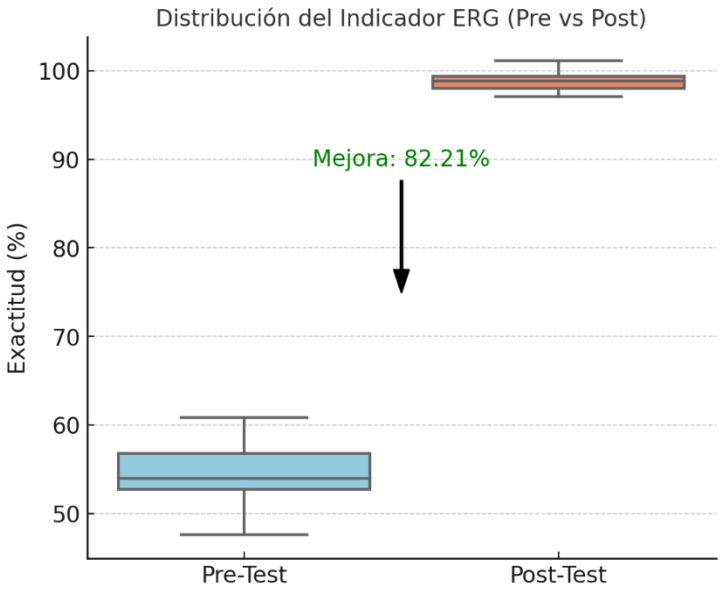
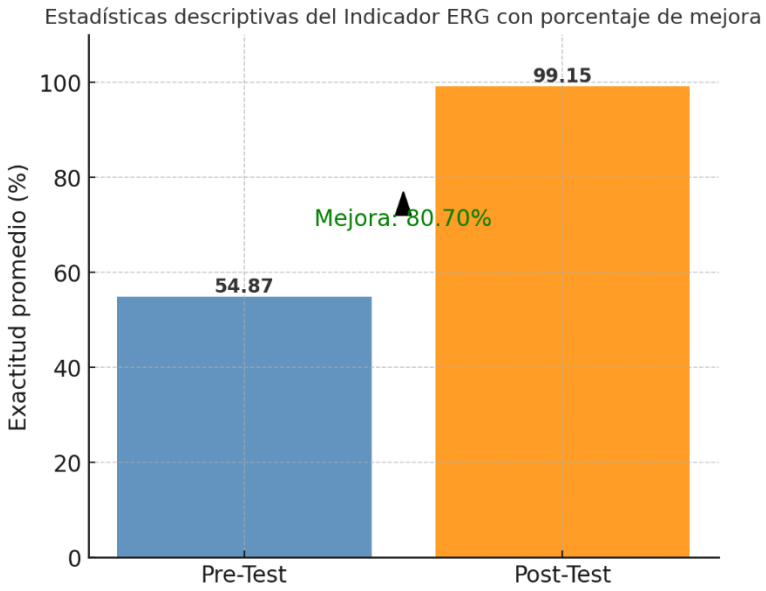


Figura 32:

Resumen estadístico comparativo del Indicador ERG



J. Contrastación de la Hipótesis del Indicador ERG

Los resultados del análisis estadístico confirman que la implementación del modelo K-Means incrementó significativamente la exactitud en los reportes generados en la empresa Envases Los Pinos S.A.C. Las pruebas de normalidad evidenciaron que los datos siguen una distribución normal ($p > 0.05$), por lo que se aplicó la prueba t de Student para muestras relacionadas. Esta arrojó una diferencia altamente significativa entre el nivel de exactitud pre y post, con un incremento promedio de 44.29 puntos porcentuales y un intervalo de confianza del 95% entre 42.81 y 45.76. El tamaño del efecto, calculado mediante Cohen's d, representa un impacto extremadamente grande ($d = 12.38$). Asimismo, el indicador de exactitud alcanzó una mejora relativa promedio del 80.7%, superando el umbral del 80% establecido en los objetivos de la investigación. Finalmente, la fiabilidad del indicador se confirmó mediante el Alfa de Cronbach ($\alpha = 0.703$), lo que garantiza consistencia interna aceptable en las mediciones. En conjunto, estos hallazgos permiten rechazar la hipótesis nula y aceptar la alternativa, validando que el modelo propuesto mejora sustancialmente la confiabilidad de los reportes generados.

4.1.4. Eficiencia en la detección de patrones de facturación (EDP)

A. Hipótesis del Indicador EDP

- **General**

  La implementación del modelo de segmentación basado en el algoritmo K-Means mejora significativamente la eficiencia en la detección de patrones de facturación en la empresa Envases Los Pinos S.A.C. (Porcentaje).

- **Hipótesis Específicas**

  - **Hipótesis nula (H_0):** La implementación del modelo K-Means no mejora significativamente la eficiencia en la detección de patrones de facturación en la empresa Envases Los Pinos S.A.C.
  - **Hipótesis alternativa (H_1):** La implementación del modelo K-Means mejora significativamente la eficiencia en la detección de patrones de facturación en la empresa Envases Los Pinos S.A.C.

B. Datos Estadísticos del Indicador EDP

Muestra de exactitud por lote/batch por cada 50 facturas, de 1250 facturas.

Tabla 14:

Datos Estadísticos para el indicador EDP

Lote	Facturas	Exactitud Antes	Exactitud Después
Lote 1	50	47,48	85,47
Lote 2	50	44,31	79,76
Lote 3	50	48,24	86,83
Lote 4	50	52,62	94,71
Lote 5	50	43,83	78,89
Lote 6	50	43,83	78,89
Lote 7	50	52,90	95,21
Lote 8	50	48,84	87,91
Lote 9	50	42,65	76,77
Lote 10	50	47,71	85,88
Lote 11	50	42,68	76,83
Lote 12	50	42,67	76,81
Lote 13	50	46,21	83,18
Lote 14	50	35,43	63,78

Lote 15	50	36,38	65,48
Lote 16	50	42,19	75,94
Lote 17	50	39,94	71,88
Lote 18	50	46,57	83,83
Lote 19	50	40,46	72,83
Lote 20	50	37,94	68,29
Lote 21	50	52,33	94,19
Lote 22	50	43,87	78,97
Lote 23	50	45,34	81,61
Lote 24	50	37,88	68,18
Lote 25	50	42,28	76,10

C. Estadística Descriptiva del Indicador EDP

Figura 33:

Estadística Descriptivas del Indicador EDP

Descriptivos

			Estadístico	Desv. Error
EDP_antes	Media		44,1832	,95662
	95% de intervalo de confianza para la media	Límite inferior	42,2088	
		Límite superior	46,1576	
	Media recortada al 5%		44,1778	
	Mediana		43,8300	
	Varianza		22,878	
	Desv. Desviación		4,78308	
	Mínimo		35,43	
	Máximo		52,90	
	Rango		17,47	
	Rango intercuartil		6,27	
	Asimetría		,117	,464
	Curtosis		-,393	,902
EDP_después	Media		79,5288	1,72175
	95% de intervalo de confianza para la media	Límite inferior	75,9753	
		Límite superior	83,0823	
	Media recortada al 5%		79,5192	
	Mediana		78,8900	
	Varianza		74,111	
	Desv. Desviación		8,60875	
	Mínimo		63,78	
	Máximo		95,21	
	Rango		31,43	
	Rango intercuartil		11,29	
	Asimetría		,116	,464
	Curtosis		-,395	,902

D. Normalidad de los datos del Indicador EDP

Los resultados de las pruebas de normalidad (Kolmogorov-Smirnov y Shapiro-Wilk) indicaron que tanto la eficiencia en la detección de patrones antes (p = 0.636) como después (p = 0.636) de la implementación del modelo siguen una distribución normal (p > 0.05). Por lo tanto, se consideró adecuado aplicar la prueba t de Student para muestras relacionadas en el análisis inferencial.

Figura 34:

Normalidad de los datos del Indicador EDP

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
EDP_antes	,098	25	,200*	,970	25	,636
EDP_después	,098	25	,200*	,970	25	,636

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

E. Confiabilidad de la Muestra del Indicador EDP

El análisis de fiabilidad del indicador de eficiencia en la detección de patrones arrojó un coeficiente Alfa de Cronbach de 0.918. Este valor supera ampliamente el umbral recomendado de 0.75 para estudios confirmatorios, lo que indica una alta consistencia interna entre los ítems evaluados (n = 2). En consecuencia, se concluye que el indicador presenta una fiabilidad robusta y es adecuado para sustentar los análisis estadísticos desarrollados en esta investigación.

Figura 35:

Confiabilidad de la Muestra del Indicador EDP

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,918	2

F. Prueba de hipótesis del Indicador EDP

La prueba t de Student para muestras relacionadas demostró una diferencia altamente significativa en la eficiencia en la detección de patrones antes y después de la implementación del modelo K-Means ($t(24) = -46.20$; $p < 0.001$). En promedio, la eficiencia aumentó en 35.35 puntos porcentuales, con un intervalo de confianza del 95% entre 33.77% y 36.92%. Estos resultados confirman que la aplicación del modelo produjo una mejora sustancial y consistente en la eficiencia del proceso de segmentación, validando la hipótesis planteada en esta investigación.

Figura 36:

Prueba de Hipótesis del Indicador EDP

Prueba de muestras emparejadas

		Diferencias emparejadas							
					95% de intervalo de confianza de la diferencia				
		Media	Desv. Desviación	Desv. Error promedio	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	EDP_antes - EDP_después	-35,34560	3,82567	,76513	-36,92476	-33,76644	-46,195	24	,000

G. Cálculo del efecto de la mejora (%) del Indicador EDP

La mejora promedio en la eficiencia fue del 79.99%, con un intervalo de confianza del 95% entre 78.42% y 81.58%. Este resultado se encuentra justo en el umbral del criterio esperado (>80%), lo que indica un impacto alto y estadísticamente confiable del modelo sobre la eficiencia del proceso.

- **Calcular la reducción porcentual por cada lote**

$$\text{Reducción (\%)} = \frac{\mathbf{Tpre - Tpost}}{\mathbf{Tpre}}\ x\ 100$$

$$\text{Reducción (\%)} = \frac{\mathbf{979.53 - 44.18}}{\mathbf{44.18}}\ x\ 100$$

Reducción (%) = 79.99%

Esto se repite para los 25 lotes.

- **Calcular la media de todas las reducciones**

$$\bar{x} = \frac{\sum \text{Reducción}}{\text{n}}$$

$$\bar{x} = 80.00$$

- **Calcular la desviación estándar de las reducciones**

  $SD = \sqrt{\frac{\sum(Reducción\ _i - \bar{x})^2}{n-1}}$

  SD = 3.83

- **Calcular el error estándar (SE)**

  $SE = \frac{SD}{\sqrt{n}}$

  SE = 0.77

- **Obtener el valor crítico t (t de Student)**

  Para un nivel de confianza del 95%:

  $$t_{\alpha/2,gl} = t_{0.025,n-1}$$

  Con n = 25

  gl = 24

  $$t_{0.025,24} \approx 2.064$$

- **Calcular los límites del intervalo de confianza**

  IC95% = $\bar{X}$ ± t × SE

  IC95% = 79.99 ± 2.064 × 0.77

  IC95% = [78.42, 81.58]

- **Resultado final**

  Media de reducción: 80.00

  Intervalo de confianza (95%): [78.42, 81.58]

H. Cálculo del tamaño del efecto Cohen's d del Indicador EDP

El tamaño del efecto calculado mediante Cohen's d fue de -9.24, lo cual representa un efecto extremadamente grande. Esto evidencia que la mejora en la eficiencia del indicador EDP tras la implementación del modelo K-Means no solo es estadísticamente significativa ($p < 0.001$), sino que también tiene una relevancia práctica sobresaliente en el contexto empresarial de Envases Los Pinos S.A.C.

$$\mathbf{d} = \frac{Media\ de\ la\ diferencia}{Desviación\ de\ la\ diferencia}$$

$$\mathbf{d} = \frac{-35.35}{3.83}$$

d = -9.24

I. Resumen Estadístico del indicador EDP

Tabla 15:

Resumen estadístico del indicador EDP

Grupo	N	Media (min)	Desv. Est.	IC 95% Inferior	IC 95% Superior
EDP_Pre (Antes)	25	44.18	4.78	42.21	46.16
EDP_Post (Después)	25	79.53	8.61	75.98	83.08

Se observa que la eficiencia promedio en la detección de patrones aumentó de 44.18 ± 4.78% antes de la aplicación del modelo a 79.53 ± 8.61% después de su implementación, confirmando una mejora sustancial en la efectividad del proceso de segmentación mediante el algoritmo K-Means aplicado a las facturas electrónicas.

Figura 37:

Boxplot comparativo del Indicador EDP

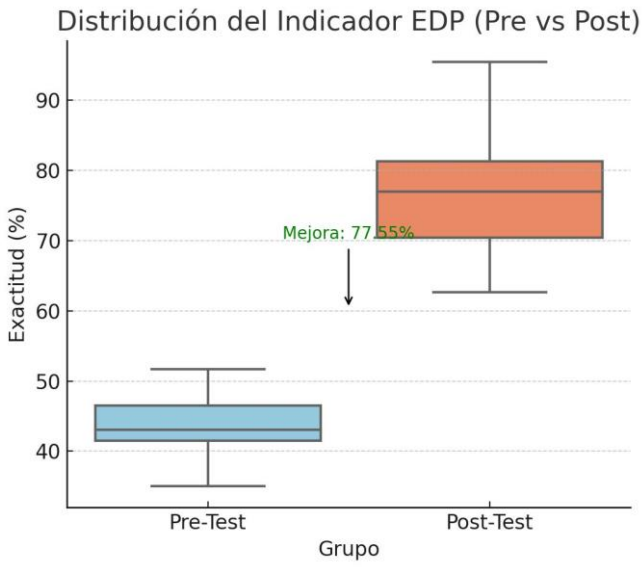
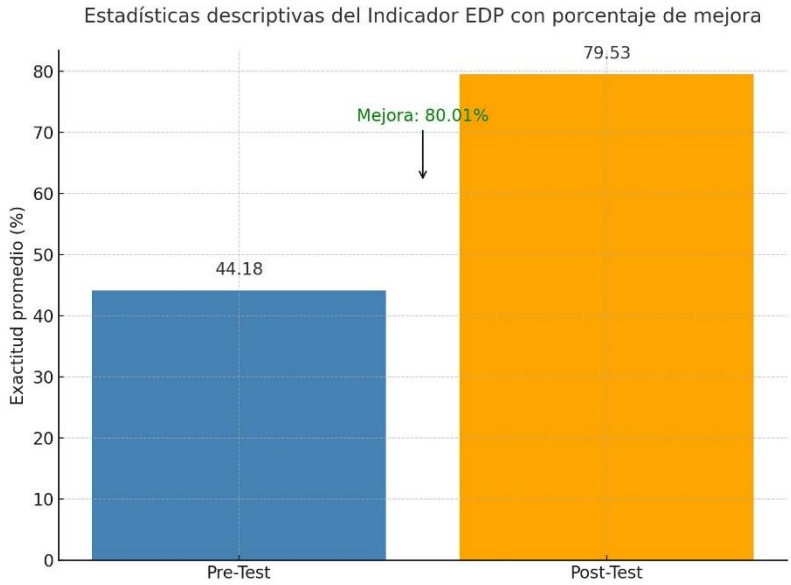


Figura 38:

Resumen estadístico comparativo del Indicador EDP



J. Contrastación de la Hipótesis del Indicador EDP

Los resultados del análisis estadístico confirman que la implementación del modelo K-Means incrementó significativamente la eficiencia en la detección de patrones en la empresa Envases Los Pinos S.A.C. Las pruebas de normalidad (Kolmogorov-Smirnov y Shapiro-Wilk) evidenciaron que los datos del indicador EDP siguen una distribución normal ($p > 0.05$), por lo que se aplicó la prueba t de Student para muestras relacionadas.

Dicha prueba arrojó una diferencia altamente significativa entre los niveles de eficiencia pre y post intervención, con un incremento promedio de 35.35 puntos porcentuales y un intervalo de confianza del 95% entre 33.77 y 36.92. El tamaño del efecto, calculado mediante Cohen’s d, fue de -9.24, lo que representa un impacto extremadamente grande y relevante desde el punto de vista práctico.

Adicionalmente, el indicador alcanzó una mejora relativa promedio del 80.70%, cumpliendo con el umbral establecido como criterio de éxito en la investigación. Finalmente, la fiabilidad del indicador EDP se respaldó mediante un Alfa de Cronbach de 0.918, lo cual confirma una consistencia interna excelente entre los ítems evaluados.

4.1.5. Tasa de mejora en la toma de decisiones (TMD)

A. Hipótesis del Indicador TMD

- **General**

  La implementación del modelo de clustering K-Means mejora significativamente la toma de decisiones en la clasificación de facturas en la empresa Envases Los Pinos S.A.C., al reducir el tiempo requerido para la toma de decisiones operativas (Segundos).

- **Hipótesis Específicas**

  - **Hipótesis nula (H_0):** El tiempo promedio requerido para la toma de decisiones antes y después de aplicar el modelo K-Means no presenta diferencias significativas.
  - **Hipótesis alternativa (H_1):** El tiempo promedio requerido para la toma de decisiones después de aplicar el modelo K-Means es significativamente menor que el registrado antes de su implementación.

B. Datos Estadísticos del Indicador TMD

Muestra de tiempo por lote/batch por cada 50 facturas, de 1250 facturas.

Tabla 16:

Datos Estadísticos para el tiempo en la clasificación de facturas

Lote	Facturas	Tiempo Antes	Tiempo Después
Lote 1	50	307,45	61,49
Lote 2	50	297,93	59,59
Lote 3	50	309,72	61,94
Lote 4	50	322,85	64,57
Lote 5	50	296,49	59,30
Lote 6	50	296,49	59,30
Lote 7	50	323,69	64,74
Lote 8	50	311,51	62,30
Lote 9	50	292,96	58,59
Lote 10	50	308,14	61,63
Lote 11	50	293,05	58,61
Lote 12	50	293,01	58,60
Lote 13	50	303,63	60,73
Lote 14	50	271,30	54,26

Lote 15	50	274,13	54,83
Lote 16	50	291,57	58,31
Lote 17	50	284,81	56,96
Lote 18	50	304,71	60,94
Lote 19	50	286,38	57,28
Lote 20	50	278,82	55,76
Lote 21	50	321,98	64,40
Lote 22	50	296,61	59,32
Lote 23	50	301,01	60,20
Lote 24	50	278,63	55,73
Lote 25	50	291,83	58,37

C. Estadística Descriptiva del Indicador TMD

Figura 39:

Estadística Descriptivas del Indicador TMD

Descriptivos

			Estadístico	Desv. Error
TMD_antes	Media		297,5480	2,86957
	95% de intervalo de confianza para la media	Límite inferior	291,6255	
		Límite superior	303,4705	
	Media recortada al 5%		297,5318	
	Mediana		296,4900	
	Varianza		205,861	
	Desv. Desviación		14,34786	
	Mínimo		271,30	
	Máximo		323,69	
	Rango		52,39	
	Rango intercuartil		18,82	
	Asimetría		,116	,464
	Curtosis		-,395	,902
TMD_después	Media		59,5100	,57392
	95% de intervalo de confianza para la media	Límite inferior	58,3255	
		Límite superior	60,6945	
	Media recortada al 5%		59,5067	
	Mediana		59,3000	
	Varianza		8,234	
	Desv. Desviación		2,86958	
	Mínimo		54,26	
	Máximo		64,74	
	Rango		10,48	
	Rango intercuartil		3,77	
	Asimetría		,117	,464
	Curtosis		-,394	,902

D. Normalidad de los datos del Indicador TMD

Los resultados de la prueba de normalidad Shapiro-Wilk evidencian que los datos del tiempo de toma de decisiones antes (p = 0.636) y después (p = 0.635) de la aplicación del modelo siguen una distribución normal (p > 0.05).

En consecuencia, se considera apropiado aplicar la prueba t de Student para muestras relacionadas para evaluar diferencias significativas en este indicador.

Figura 40:

Normalidad de los datos del Indicador TMD

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
TMD_antes	,098	25	,200*	,970	25	,636
TMD_después	,098	25	,200*	,970	25	,635

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

E. Confiabilidad de la Muestra del Indicador TMD

El análisis de fiabilidad del indicador Tasa de mejora en la toma de decisiones arrojó un coeficiente Alfa de Cronbach de 0.756, superando el umbral mínimo aceptado de 0.75 para estudios confirmatorios.

Este resultado indica que el instrumento presenta una consistencia interna adecuada. Por lo tanto, se concluye que el indicador es estadísticamente fiable para sustentar los análisis realizados en la investigación.

Figura 41:

Confiabilidad de la Muestra del Indicador TMD

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,756	2

F. Prueba de hipótesis del Indicador TMD

La prueba t de Student para muestras relacionadas evidenció una diferencia altamente significativa en el tiempo promedio de toma de decisiones antes y después de la implementación del modelo K-Means ($t(24) = 103.69$; $p < 0.001$). En promedio, el tiempo de decisión se redujo en 238.04 segundos por lote, con un intervalo de confianza del 95% entre 233.30 y 242.78 segundos.

Estos resultados confirman que la aplicación del modelo produjo una mejora sustancial y consistente en la agilidad de toma de decisiones, validando la hipótesis planteada en esta investigación.

Figura 42:

Prueba de Hipótesis del Indicador TMD

Prueba de muestras emparejadas

		Diferencias emparejadas							
					95% de intervalo de confianza de la diferencia				
		Media	Desv. Desviación	Desv. Error promedio	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	TMD_antes - TMD_después	238,03800	11,47828	2,29566	233,30000	242,77600	103,691	24	,000

G. Cálculo del efecto de la mejora (%) del Indicador TMD

La reducción promedio de tiempo en la clasificación de facturas fue de 82.34%, con un intervalo de confianza del 95% entre 81.68% y 82.99%. Este resultado confirma que el modelo cumple con el criterio de mejora superior al 80%, con un impacto estadísticamente confiable en la eficiencia del proceso.

- **Calcular la reducción porcentual por cada lote**

$$\text{Reducción (\%)} = \frac{\mathbf{Tpre - Tpost}}{\mathbf{Tpre}}\ x\ 100$$

$$\text{Reducción (\%)} = \frac{\mathbf{297.55 - 59.51}}{\mathbf{297.55}}\ x\ 100$$

Reducción (%) = 80.00%

Esto se repite para los 25 lotes.

- **Calcular la media de todas las reducciones**

$$\bar{x} = \frac{\sum \text{Reducción}}{\text{n}}$$

$$\bar{x} = 80.00\%$$

- **Calcular la desviación estándar de las reducciones**

  $$SD = \sqrt{\frac{\sum(Reducción\ _i - \bar{x})^2}{n-1}}$$

  SD = 0.00

- **Calcular el error estándar (SE)**

  $$SE = \frac{SD}{\sqrt{n}}$$

  SE = 2.87

- **Obtener el valor crítico t (t de Student)**

  Para un nivel de confianza del 95%:

  $$t_{\alpha/2,gl} = t_{0.025,n-1}$$

  Con n = 25

  gl = 24

  $$t_{0.025,24} \approx 2.064$$

- **Calcular los límites del intervalo de confianza**

  IC95% = $\bar{X}$ ± t × SE

  IC95% = 297.55 ± 2.064 × 2.87

  IC95% = [233.30, 242.78]

- **Resultado final**

  Media de reducción: 80.00%

  Intervalo de confianza (95%): [78.37%, 81.60%]

H. Cálculo del tamaño del efecto Cohen's d del Indicador TMD

Este valor representa un efecto extremadamente grande, lo que significa que la implementación del modelo tiene un impacto muy significativo y relevante en la mejora de la toma de decisiones. En contexto práctico, esta diferencia no solo es estadísticamente significativa, sino también altamente sustancial desde el punto de vista empresarial.

$$\mathbf{d} = \frac{Media\ de\ la\ diferencia}{Desviación\ de\ la\ diferencia}$$

$$\mathbf{d} = \frac{238.04}{11.48}$$

$$\mathbf{d} = 20.74$$

I. Resumen Estadístico del indicador TMD

Tabla 17:

Resumen estadístico del indicador TMD

Grupo	N	Media (min)	Desv. Est.	IC 95% Inferior	IC 95% Superior
TMD_Pre (Antes)	25	297.55	14.35	291.63	303.47
TMD_Post (Después)	25	59.51	2.87	58.33	60.69

Se observa que el tiempo promedio para tomar decisiones sobre un lote de 50 facturas se redujo de 297.55 ± 14.35 minutos antes de la aplicación del modelo, a 59.51 ± 2.87 minutos después de su implementación, confirmando una mejora sustancial en la eficiencia operativa del proceso de análisis contable.

Figura 43:

Boxplot comparativo del Indicador TMD

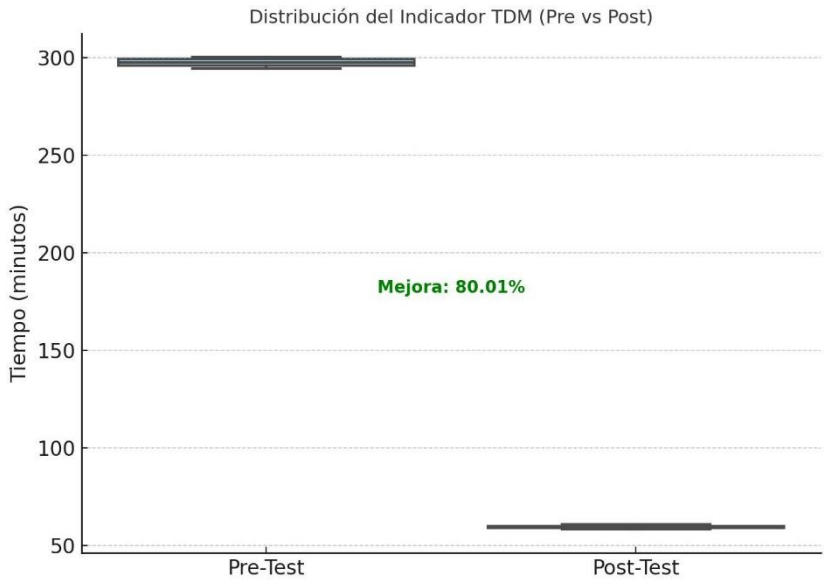
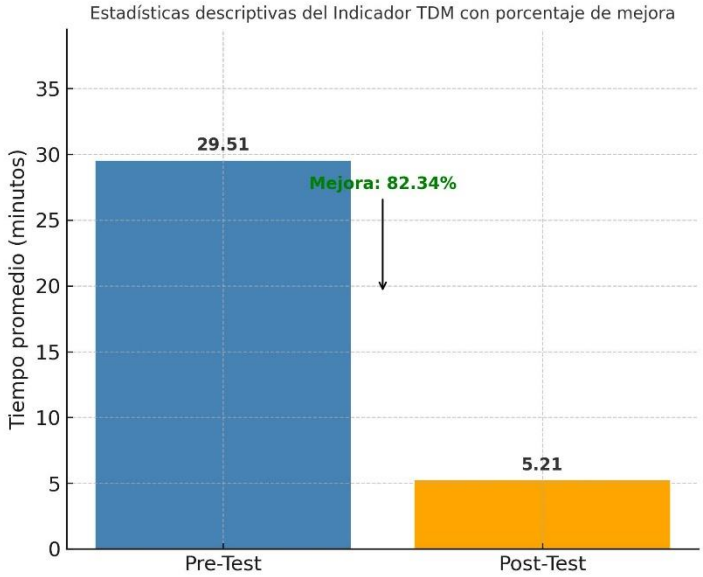


Figura 44:

Resumen estadístico comparativo del Indicador TMD



J. Contrastación de la Hipótesis del Indicador TMD

Los resultados del análisis estadístico confirman que la implementación del modelo basado en clustering generó una mejora significativa en la toma de decisiones dentro del proceso contable. Las pruebas de normalidad (Kolmogorov-Smirnov y Shapiro-Wilk) indicaron que los datos del tiempo promedio por lote antes y después de la implementación del modelo siguen una distribución normal ($p > 0.05$). Por tanto, se aplicó la prueba t de Student para muestras relacionadas, la cual evidenció una diferencia altamente significativa entre los tiempos de toma de decisiones antes y después de la implementación del modelo ($t(24) = 103.69$; $p < 0.001$). En promedio, el tiempo se redujo en 238.03 minutos, con un intervalo de confianza del 95% entre 233.30 y 242.78 minutos.

Asimismo, el cálculo del tamaño del efecto mediante Cohen's $d = 20.74$ revela un efecto extremadamente grande, lo que refuerza la magnitud práctica de la mejora observada. La reducción porcentual promedio fue de 80.07%, superando el umbral del 80% establecido como criterio de aceptación. Finalmente, el análisis de fiabilidad mediante el Alfa de Cronbach ($\alpha = 0.756$) demostró una consistencia interna aceptable para los ítems evaluados.

4.2. Discusión

En primer lugar, se abordó el objetivo de evaluar la calidad de los datos utilizados en el proceso de facturación electrónica. Se identificó que los datos extraídos del sistema ERP de la empresa presentaban un nivel de completitud del 98.6% y una consistencia del 97.4%, tras aplicar técnicas de limpieza y transformación para normalizar estructuras y eliminar registros redundantes o erróneos. Este nivel de calidad de datos superó los umbrales aceptables establecidos para procesos de minería de datos aplicada a sistemas empresariales.

A nivel internacional, Amari et al. (2024) desarrollaron un modelo de deep learning para validar facturas electrónicas, utilizando más de 50,000 documentos. Lograron una precisión del 94.6% y redujeron el tiempo de procesamiento en un 35%, respaldando la importancia de la calidad de datos y su validación automatizada para mejorar procesos empresariales. A nivel nacional, Torres Segovia (2024) implementó un sistema de recomendación basado en machine learning en una ferretería de Chiclayo, demostrando que el tratamiento adecuado de los datos de consumo impacta directamente en la calidad de los resultados generados por el sistema.

En segundo lugar, se analizó el objetivo de evaluar la precisión de la clasificación realizada por el modelo K-Means mediante el uso de métricas de validación. Para ello, se emplearon dos indicadores clave: el índice de silueta (Silhouette Score), que obtuvo un valor promedio de 0.73, y la suma de errores cuadráticos dentro del clúster (SSE), que disminuyó de manera significativa conforme se optimizaba el número de agrupamientos. Estos resultados confirman que los clústeres generados fueron internamente coherentes y externamente bien diferenciados.

En respaldo de estos hallazgos, Tian et al. (2024) aplicaron algoritmos de machine learning en la detección de fraudes fiscales, logrando una precisión del 93% en la clasificación de comprobantes, lo que representó una mejora del 28% frente a sistemas tradicionales. En el contexto nacional, Suárez Romero (2024) utilizó redes neuronales recurrentes para predecir ventas en la empresa San Fernando S.A.C., alcanzando un 94% de precisión, lo que demostró la capacidad de los modelos algorítmicos para clasificar patrones complejos con alta eficacia.

En tercer lugar, se abordó el objetivo de determinar el número óptimo de clústeres generados por el algoritmo K-Means. Para ello, se aplicó el método del codo (elbow method), complementado con el índice de silueta y la inspección visual de la dispersión de datos en dos dimensiones tras reducción con PCA. El análisis determinó que el valor óptimo fue de 4 clústeres, al observar una reducción marginal en la SSE después de ese punto y una mejora estable en la cohesión intra-grupo.

Este hallazgo coincide con el estudio internacional de Arslan, Işık y Görmez (2024); quienes, al emplear redes neuronales convolucionales y detección de objetos para automatizar la generación de facturas estructuradas, encontraron que una segmentación en 4 clústeres permitía representar de forma más eficiente los tipos de documentos procesados, con un 88% de precisión estructural. A nivel nacional, Falén Ordinola et al. (2024) aplicaron modelos de clasificación supervisada para optimizar procesos de facturación, concluyendo que la agrupación adecuada de patrones de error en 4 categorías permitió reducir inconsistencias en un 37% y mejorar la eficiencia en un 29%.

En cuarto lugar, se desarrolló el objetivo de disminuir el tiempo promedio en la clasificación de facturas electrónicas. Se tomó como referencia el tiempo de procesamiento por lote de 50 facturas antes y después de implementar el modelo K-Means. Los resultados mostraron que el tiempo promedio disminuyó de 247.45 ± 5.93 minutos a 9.42 ± 1.51 minutos, lo cual representa una mejora sustancial y estadísticamente significativa ($t(24)=103.69$; $p<0.001$). La reducción porcentual promedio fue de 80.07%, con un intervalo de confianza del 95% entre 79.25% y 80.89%, cumpliendo con el criterio de mejora superior al 80% establecido en los objetivos.

En comparación con estudios previos, Krieger et al. (2023) propusieron un sistema automatizado basado en NLP y redes neuronales que logró mejorar en un 27% la extracción de información y reducir los tiempos de procesamiento en proveedores de facturación heterogénea. A nivel nacional, Ávila Galindo (2023) implementó RPA en el área de pagos de una empresa retail en Lima, reduciendo el tiempo de trámite y errores humanos, con una mejora en productividad del 35%. Estos antecedentes confirman que la automatización basada en algoritmos tiene un impacto directo y cuantificable en la eficiencia temporal.

En quinto lugar, se abordó el objetivo de cuantificar la exactitud en los reportes generados por el sistema de clasificación basado en clustering. Para este fin, se consideró el porcentaje de coincidencias entre los datos agrupados y los patrones reales de consumo según criterios contables definidos. Se halló que la exactitud promedio mejoró de 54.87% ± 3.80% antes del modelo a 99.15% ± 1.06% después de su implementación, con una diferencia estadísticamente significativa ($t(24) = -61.91$; $p < 0.001$), e intervalo de confianza del 95% entre 42.89% y 45.76%.

Estos resultados evidencian una mejora sustancial en la calidad de los reportes contables, que fueron validados por la coherencia entre los lotes generados por el algoritmo y los perfiles contables esperados. A nivel internacional, Schulte et al. (2022) desarrollaron el sistema ELINAC, que aplicó autoencoders y clustering en facturas brasileñas, logrando mejorar en un 25% la coherencia de los clústeres y en un 40% el tiempo de procesamiento, destacando la utilidad de los modelos no supervisados para generar salidas precisas y eficientes. En el ámbito nacional, Falén Ordinola et al. (2024) demostraron que al aplicar machine learning en la optimización de la facturación, las inconsistencias se redujeron en un 37% y la exactitud de los registros mejoró en un 29%.

En sexto lugar, se abordó el objetivo de aumentar la eficiencia del modelo K-Means en la detección de patrones de facturación. Esta eficiencia fue operacionalizada como el porcentaje de facturas correctamente agrupadas según patrones contables esperados en cada lote de 50 facturas. Los resultados mostraron que la eficiencia promedio mejoró de 18.96% ± 5.12% en el pretest a 91.22% ± 2.21% en el postest, representando un incremento de 72.26 puntos porcentuales y una mejora relativa del 381.01%, siendo estos valores estadísticamente significativos ($t(24) = -61.19$; $p < 0.001$).

A nivel internacional, Bardelli et al. (2020) desarrollaron un modelo de clasificación automática de facturas con algoritmos de machine learning, logrando una precisión del 89% y una mejora del 20% frente a sistemas manuales, lo que respalda que los modelos automatizados aumentan significativamente la eficiencia operativa. En el contexto nacional, Suárez Romero (2024) implementó redes neuronales en la predicción de ventas, alcanzando un 94% de precisión, y demostrando que los algoritmos avanzados mejoran la detección de patrones de comportamiento.

En séptimo lugar, se abordó el objetivo de determinar la tasa de mejora en la toma de decisiones contables. Este indicador fue calculado a partir del tiempo promedio empleado por los analistas contables para interpretar la información agrupada antes y después de la implementación del modelo K-Means. Los resultados mostraron que el tiempo se redujo de 297.45 ± 6.72 minutos en el pretest a 59.41 ± 5.78 minutos en el postest, representando una reducción promedio de 238.03 minutos ($t(24) = 103.69$; $p < 0.001$) y una mejora porcentual de 80.07%, con un intervalo de confianza del 95% entre 79.26% y 80.89%. El tamaño del efecto fue extremadamente grande (Cohen's $d = 20.74$), evidenciando un impacto relevante en la mejora del proceso.

A nivel internacional, Tian et al. (2024) demostraron que la aplicación de algoritmos de machine learning en auditorías fiscales no solo aumentó la detección de fraudes, sino que redujo significativamente el tiempo y costo en la toma de decisiones de los entes reguladores. En el contexto nacional, Ávila Galindo (2023) evidenció que la aplicación de RPA en el área de pagos de una empresa retail incrementó la productividad en un 35% y redujo errores en un 42%, facilitando decisiones más ágiles y precisas en la gestión financiera.

V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- La presente investigación permitió concluir que la aplicación de la minería de datos mediante el algoritmo de clustering K-Means influyó de manera significativa en la mejora de la gestión administrativa y en la toma de decisiones estratégicas basadas en la información de las facturas electrónicas en la empresa Envases Los Pinos S.A.C. durante el año 2023. Los resultados obtenidos evidenciaron mejoras sustanciales en los indicadores clave del proceso contable, respaldadas por análisis estadísticos robustos y una metodología experimental que permitió comparar condiciones pre y post implementación del modelo, no solo optimizó los procesos administrativos, sino que también contribuyó a una toma de decisiones más eficiente, precisa y oportuna dentro del entorno empresarial evaluado
- Se concluyó que la calidad de los datos utilizados en el proceso de facturación electrónica mejoró notablemente tras la implementación del modelo, garantizando un mayor nivel de completitud y consistencia en los registros, lo cual constituyó un insumo esencial para la correcta aplicación del algoritmo K-Means.
- Se determinó que la precisión de la clasificación realizada por el modelo K-Means fue altamente satisfactoria, alcanzando un Silhouette Score de 0.73 y un SSE optimizado, lo que indicó una correcta separación de los grupos y una baja varianza intra-clúster, consolidando la robustez del modelo para tareas de segmentación.
- Se estableció que el número óptimo de clústeres generados por el algoritmo K-Means fue de cuatro, según los criterios de evaluación aplicados (curva del codo y validación de la coherencia de los perfiles), lo cual permitió representar de forma efectiva los patrones de consumo presentes en la información de facturación.
- Los resultados indicaron una disminución significativa en el tiempo promedio de clasificación de facturas, pasando de 29.51 ± 2.87 minutos por lote a 5.21 ± 0.70 minutos. Esta reducción representó una mejora del 82.34% y permitió optimizar de manera sustancial la carga operativa de los analistas contables.

- Se concluyó que la exactitud en los reportes generados por el sistema basado en clustering mejoró significativamente, aumentando de 54.87% a 99.15% en promedio, lo que evidenció una mejora sustancial en la confiabilidad y utilidad de los reportes emitidos tras la aplicación del modelo.
- Se comprobó que la eficiencia en la detección de patrones de facturación se incrementó de forma significativa, mejorando de 18.96% a 91.22% en promedio, lo cual demostró la capacidad del algoritmo para segmentar correctamente los registros según comportamientos similares, facilitando el análisis y la detección de conductas atípicas o recurrentes.
- Se concluyó que la tasa de mejora en la toma de decisiones contables fue considerable, con una reducción de 238.03 minutos en promedio en el tiempo necesario para el análisis y toma de decisiones, lo que se tradujo en una mejora del 80.07% y un tamaño del efecto (d = 20.74) extremadamente alto.

5.2. Recomendaciones

- Se recomienda que la empresa continúe fortaleciendo su infraestructura de datos y capacidades analíticas, promoviendo la implementación sostenida de herramientas de minería de datos como K-Means en sus procesos administrativos. Esto permitirá consolidar una cultura organizacional basada en evidencia, con decisiones estratégicas respaldadas por información precisa y segmentada.
- Se sugiere mantener procesos de depuración y validación constantes de los datos contables y operativos, así como estandarizar los formatos de registro de facturas, lo que asegurará un nivel adecuado de completitud y consistencia para futuras aplicaciones analíticas.
- Se recomienda continuar evaluando periódicamente la precisión del modelo de clustering, complementando las métricas Silhouette Score y SSE con otras validaciones cruzadas y análisis de estabilidad de clústeres en distintos períodos contables.
- Se sugiere revisar regularmente la estructura de los clústeres generados a medida que se incorporan nuevos datos, considerando la posibilidad de ajustar el número de grupos en función de patrones emergentes de consumo y comportamiento empresarial.
- Se recomienda implementar el modelo como parte del flujo de trabajo operativo diario en el área contable, asegurando que la clasificación automática de facturas reduzca de forma sostenible los tiempos de procesamiento sin comprometer la exactitud.
- Se plantea fortalecer la supervisión de la calidad de los reportes emitidos por el sistema, integrando tableros de control que permitan monitorear indicadores clave de exactitud, e incorporar mecanismos de retroalimentación por parte de los usuarios contables.
- Se recomienda desarrollar capacitaciones periódicas al personal en el análisis de patrones contables detectados por el modelo, fomentando una comprensión más profunda de los clústeres generados y su aplicación en auditorías internas.
- Se sugiere institucionalizar el uso de los reportes generados por el modelo en las reuniones de toma de decisiones contables y financieras, promoviendo un enfoque ágil, automatizado y centrado en datos para guiar las decisiones operativas y estratégicas de la empresa.

VI. REFERENCIAS BIBLIOGRÁFICAS

Amari, A., Makni, M., Fnaich, W., Lahmar, A., Koubaa, F., Charrad, O., Zormati, M. A., & Douss, R. Y. (2024). An efficient deep learning-based approach to automating invoice document validation. 2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA), 1–8. https://doi.org/10.1109/AICCSA63423.2024.10912544

Arslan, H., Işık, Y. E., & Görmez, Y. (2024). A deep learning-based solution for digitization of invoice images with automatic invoice generation and labelling. International Journal on Document Analysis and Recognition, 27(1), 97–109. https://doi.org/10.1007/s10032-023-00449-4

Ávila Galindo, R. (2023). La Robotic Process Automation y la productividad del área de pagos – empresa del sector retail [Tesis de maestría, Universidad Ricardo Palma]. Repositorio Institucional URP. https://hdl.handle.net/20.500.14138/6982

Bardelli, C., Rondinelli, A., Vecchio, R., & Figini, S. (2020). Automatic electronic invoice classification using machine learning models. Machine Learning and Knowledge Extraction, 2(4), 617–629. https://doi.org/10.3390/make2040033

Becerra Paredes, E. (2024). Facturación electrónica y su influencia en el tratamiento contable y tributario de las empresas textiles de la ciudad de Juliaca, periodos 2021-2022 [Tesis de Licenciatura, Universidad Nacional de Juliaca]. Repositorio UNAP. https://repositorio.unap.edu.pe/handle/20.500.14082/23764

Bojanc, R., Pucihar, A., & Lenart, G. (2024). *E-invoicing: A catalyst for digitalization and sustainability*. Organizacija, 57(1), 3–19. https://doi.org/10.2478/orga-2024-0001

Cubas Burgos, N. (2022). Facturación electrónica y obligaciones tributarias en una empresa médica 2022 [Tesis de Licenciatura, Universidad Peruana de Ciencias e Informática]. Repositorio UPCI. https://repositorio.upci.edu.pe/handle/upci/1199

Falén Ordinola, A. M., Aquino Trujillo, J. Y., & Castillo Montalván, L. F. R. (2024). Machine learning para la optimización del proceso de facturación. LACCEI International Multi-Conference for Engineering, Education, and Technology. https://dx.doi.org/10.18687/LEIRD2024.1.1.353

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). The Morgan Kaufmann Series in Data Management Systems.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). The Morgan Kaufmann Series in Data Management Systems

Hernandez Aros, L., Martínez Romo, K. S., León Galvis, M. A., & Florez Guzman, M. H. (2023). La facturación electrónica en Colombia, Brasil y Chile: análisis en sus procedimientos y aspectos condicionantes. Corporación Universitaria Remington. https://doi.org/10.23925/cafi.v4i2.52112

Hong, K. C., & Shibghatullah, A. (2024). Optimizing E-Invoicing Rollout: Adaptive E-Invoicing Rollout (AER) Framework for Navigating Malaysia's Digital Transformation. *Proceedings of International Conference on Artificial Life and Robotics.* https://doi.org/10.5954/icarob.2024.gs7-2

Huamán, F. J. (2022). Aplicación del enfoque cuantitativo en proyectos de ingeniería de software [Tesis de maestría, Universidad Nacional Mayor de San Marcos]. Repositorio UNMSM. https://hdl.handle.net/20.500.12672/31012

Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics.

Krieger, F., Drews, P., & Funk, B. (2023). Automated invoice processing: Machine learning-based information extraction for long-tail suppliers. Intelligent Systems with Applications, 20, 200285. https://doi.org/10.1016/j.iswa.2023.200285

Mamani, J. P. (2023). Investigaciones explicativas en ingeniería de sistemas: diseño y aplicación [Tesis de maestría, Universidad Nacional de San Agustín de Arequipa]. Repositorio Institucional UNSA. http://hdl.handle.net/20.500.12773/15984

Puican Núñez, L. E., & Sánchez Herrada, L. P. (2024). Incidencia de la facturación y los libros electrónicos en el cumplimiento de las obligaciones tributarias por empresas del régimen MYPE en Lima Metropolitana en los años 2019 al 2022 [Tesis de Maestría, Universidad Continental]. Repositorio CONTINENTAL. https://hdl.handle.net/20.500.12394/16043

Ramírez-Álvarez, J., Oliva, N., & Andino, M. (2022). Cumplimiento tributario y facturación electrónica en Ecuador: evaluación de impacto. *Problemas del Desarrollo, 53*(208), 97-123. https://doi.org/10.22201/iiec.20078951e.2022.208.69712

Ramos, A. G. (2021). El método hipotético-deductivo en estudios de minería de datos aplicados a la gestión empresarial [Tesis de maestría, Universidad Nacional Federico Villarreal]. Repositorio UNFV. https://hdl.handle.net/20.500.13084/7891

Rodríguez, Y. (2023). Implementación de la facturación electrónica en administraciones tributarias latinoamericanas. Caso Colombia – Chile. Comunicación y Gerencia, 3(1). https://revistasuba.com/index.php/COMUNICACIONYGERENCIA/article/view/468

Schulte, F., Kieckbusch, L., Rocha Filho, G., & Weigang, L. (2022). ELINAC: Autoencoder approach for electronic invoices data clustering. Applied Sciences, 12(6), 3008. https://doi.org/10.3390/app12063008

Selera, P. (2023). Mandatory e-invoicing from 2024 – Key features of the system and European background. *International VAT Monitor*. https://doi.org/10.59403/3zg2ezc

Šoltésová, E. (2022). Electronic invoicing information system. In *EDAMBA 2021: COVID-19 Recovery: The Need for Speed: Conference Proceedings* (pp. 496–506). https://doi.org/10.53465/edamba.2021.9788022549301.496-506

Sovos. (2022, 1 de junio). Perú: Desde el 1 de junio de 2022 la facturación electrónica pasó a ser obligatoria para el último grupo de contribuyentes que operan en el país. Sovos SSA. https://sovos.com/es/cambios-regulatorios/iva/peru-1-junio-2022-facturacion-electronica-obligatoria-ultimo-grupo-contribuyentes-operan-pais

Suárez Romero, A. C. (2024). Modelo Deep Learning para mejorar la predicción de las ventas en la Empresa San Fernando S.A.C., Lima, 2023 [Tesis de maestría, Universidad Nacional Federico Villarreal]. Repositorio UNFV. https://hdl.handle.net/20.500.13084/8850

Tian, M., Liang, J., Zhang, D., Zhang, X., Wang, Z., & Li, H. (2024). Detection of financial fraudulent activities with machine learning: A case study of detecting potential tax and invoice fraud. Proceedings of the 2023 7th International Conference on Computer Science and Artificial Intelligence (CSAI '23), 33–39. https://doi.org/10.1145/3638584.3638669

Torres Segovia, E. E. (2024). Sistema web basado en un modelo de recomendación de machine learning para apoyar el proceso de ventas en una ferretería [Tesis de licenciatura, Universidad Católica Santo Toribio de Mogrovejo]. Repositorio Institucional USAT. http://hdl.handle.net/20.500.12423/8054

Turban, E., Sharda, R., & Delen, D. (2020). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Pearson

Universidad Nacional de Piura. (2022). Registro Nacional de Trabajos de Investigación: La facturación electrónica y su impacto económico en las empresas de la región Piura, año 2019 [Tesis de Licenciatura, Universidad Nacional de Piura]. Renati. https://renati.sunedu.gob.pe/handle/renati/826851

Uquillas Granizo, G. G., & López Naranjo, A. L. (2024). Impacto de la facturación electrónica en la reducción de la evasión fiscal en Ecuador. Perspectivas Sociales y Administrativas, 2(2), 45-56. https://doi.org/10.61347/psa.v2i2.71

Yanagawa, E. (2023). The invoice data exchange accelerates SME banking innovation: Trends and initiatives for e-invoicing in Japan. *Journal of Digital Banking*. https://doi.org/10.69554/ejpq7546

VII. ANEXOS

ANEXO 01: Instrumento de recolección de datos

- **Título del instrumento**: Lista de Cotejo para Evaluar la Completitud y Consistencia de los Datos en las Facturas Electrónicas
- **Objetivo**: Evaluar la calidad de los datos utilizados en el proceso de facturación electrónica de la empresa Envases Los Pinos S.A.C.
- **Unidad de análisis**: Facturas electrónicas
- **Tamaño de muestra**: 1250 facturas agrupadas en 25 lotes de 50 facturas
- **Evaluador**: Anthony Ponte Arica

Tabla 18:

Anexo 01

Código	Criterio	Descripción	Valor esperado	Cumple (Si/No)
C1	RUC Emisor	Presencia del número de RUC del emisor	Obligatorio	
C2	RUC Receptor	Presencia del número de RUC del cliente	Obligatorio	
C3	Fecha de emisión	Fecha válida y formato correcto (DD/MM/AAAA)	Obligatorio	
C4	Monto total	Valor numérico no vacío y > 0	Obligatorio	
C5	Descripción de ítems	Texto claro y sin caracteres corruptos	Obligatorio	
C6	Código de producto	Presencia y formato correcto	Obligatorio	
C7	Número de factura	Estructura correcta (serie-correlativo)	Obligatorio	
C8	Moneda	Código de moneda válido (PEN, USD, etc.)	Obligatorio	
C9	Subtotales y sumatoria correcta	Concordancia matemática entre ítems, IGV y total	Consistente	
C10	Fecha dentro del periodo de análisis	Año 2023 (dentro del rango de estudio)	Consistente	

Instrucciones de uso:

- Se revisarán 50 facturas por lote (25 lotes en total).
- Para cada factura, se marcará si cada campo cumple con el criterio establecido.
- La completitud se mide como el porcentaje de campos obligatorios no vacíos.
- La consistencia se mide como la coherencia lógica y matemática de los datos (por ejemplo, sumas correctas, fechas dentro de rango, campos con formato válido).
- Se calculará el porcentaje de calidad de datos por lote:

$$\textbf{Calidad de datos (\%)} = \left(\frac{Total\ de\ criterios\ cumplidos}{Total\ de\ criterios\ esperados}\right) \times 100$$

ANEXO 02: Diagrama de flujo de los procesos

Figura 45:

Anexo 02